iejee

March 2025, Volume 17, Issue 2

www.iejee.com

# international electronic journal of elementary education

## Special Issue

Large Scale Assessment: Challenges and Innovations

**Editors:**

**Ummugul Bezirhan**
*Boston College*

**Murat Doğan Şahin**
*Anadolu University*

KURA group Education & Publishing

# INTERNATIONAL ELECTRONIC JOURNAL OF ELEMENTARY EDUCATION




ISSN: 1307-9298

www.iejee.com
iejee@iejee.com

KURA group
Education & Publishing

**This page is intentionally left blank.**

www.**iejee**.com

iejee

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

# Table of Contents




ISSN: 1307-9298

iejee

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

# Editorial for the Special Issue on Large Scale Assessment: Challenges and Innovations

Ummugul Bezirhan, Murat Doğan Şahin

Received : 01 March 2025
Revised : 01 March 2025
Accepted : 19 March 2025
DOI : 10.26822/iejee.2025.381

[a] Ummugul Bezirhan, TIMSS & PIRLS International Study Center, Boston College, USA.
E-mail: bezirhan@bc.edu
ORCID: https://orcid.org/0000-0002-8771-4780

[b] Murat Doğan Şahin, Faculty of Education, Anadolu University, Eskişehir, Türkiye.
E-mail: muratdogansahin@gmail.com
ORCID: https://orcid.org/0000-0002-2174-8443

## Abstract

This editorial introduces the IEJEE's Special Issue on Large Scale Assessment: Challenges and Innovations, highlighting emerging themes and methodological advancements in educational measurement. The selected studies focus on process data utilization to examine test-taker behavior, innovations in psychometric modeling for assessment, classification, and the influence of social-emotional learning on academic achievement. This editorial discusses the contributions of the included studies, their implications for future research, and the evolving role of AI, machine learning, and digital assessment technologies in shaping the future of large-scale assessments.

## Introduction

Educational testing is transforming dramatically as digital platforms, data analytics, and machine learning reshape assessment practices. These advancements provide deeper insights into test-taker behavior, enhance psychometric modeling techniques, and expand our understanding of the cognitive and non-cognitive factors influencing student achievement. As educational systems worldwide embrace digital testing and artificial intelligence (AI)-driven methodologies, researchers must navigate both the opportunities and challenges presented by these innovations.

One of the most profound shifts in large-scale assessment research is the increasing reliance on process data to capture real-time student interactions during testing. Process data allows researchers to analyze patterns of engagement, test-taking strategies, and response modifications, providing a richer picture of student performance (e.g., Bezirhan, 2021; Goldhammer et al., 2014; Ulitzsch et al., 2020; Wise 2017), insights that were largely inaccessible in paper-based assessment environments (Kane & Mislevy, 2017). Additionally, well established methodologies such as latent class analysis (LCA) and latent profile analysis (LPA), offer powerful tools for identifying unobserved subgroups of test-takers, allowing researchers to refine student classifications (Williams & Kibowski, 2016). LCA has been explored as a data-driven

alternative for setting proficiency classifications in assessments (e.g., Templin & Jiao, 2012; Binici & Cuhadar, 2022). Similarly, recent applications of LPA in educational measurement have been particularly effective in analyzing test-taking engagement (e.g., Anghel et al., 2025) and variations in student problem solving strategies (e.g, Teig, 2024) across different populations utilizing process data.

Beyond the technical innovations, the role of non-cognitive factors in student achievement has garnered attention as well. Traditional assessments have long focused on cognitive abilities and content knowledge, but emerging research highlights the importance of social-emotional learning (SEL), test-taking motivation, and engagement as key predictors of achievement (OECD, 2021). These factors not only shape student performance but also raise important considerations for fairness and equity in assessment design. This holistic approach acknowledges that academic achievement is shaped by complex interactions between content knowledge, test taking strategies, and non-cognitive factors such as perseverance, self-regulation, and social awareness (Farrington et al., 2012).

This special issue brings together a collection of studies balancing established methodologies with emerging advancements to address challenges faced in large-scale assessments. The included articles explore the intersection of process data, psychometric innovation, and non-cognitive influences on learning outcomes. By addressing both theoretical and practical implications, this issue offers fresh perspectives on how assessment research can evolve to meet the demands of contemporary education.

**Overview of the Special Issue**

As large-scale assessments continue to evolve, researchers explore novel approaches to address fundamental challenges in educational measurement. The studies featured in this special issue contribute to this growing body of research by examining innovations in process data, psychometric modelling, and non-cognitive measurement. The research presented here spans a diverse range of assessment contexts, including international assessments such as the Programme for International Student Assessment (PISA) and the Progress in International Reading Literacy Study (PIRLS), national assessments like the National Assessment of Educational Progress (NAEP), and state-level assessments such as the Iowa Assessments and the Cognitive Abilities Test (CogAT). While these studies employ distinct analytical frameworks, they collectively enhance our understanding of how large-scale assessments can be designed, analyzed, and interpreted to better support diverse student populations.

A key area of innovation in this issue is process data and student behavior analysis. The study by Ogut et al. (in this issue) examines extended time (ET) accommodations in the NAEP Grade 8 Mathematics assessment, utilizing a machine learning model (XGBoost) to identify students who may benefit from additional time. Their findings indicate that while a majority of students granted ET do not fully utilize it, nearly a quarter of students without accommodations remain actively engaged when their time expires. This study highlights the potential for predictive models to guide more equitable ET allocation policies, confirming that students with actual needs receive appropriate support. Ni et al. (in this issue) investigate response change behaviors in NAEP constructed response items, developing a novel framework that integrates automated scoring with dimensional response analysis. Their study finds that students who make substantive changes to their responses, particularly those involving conceptual modifications, are more likely to improve their scores. This research underscores the value of process data in understanding student engagement and response strategies, paving the way for more adaptive scoring and feedback mechanisms in digital assessments. Kara (in this issue) explores test-taking disengagement in PISA 2022, using LPA to classify students based on response time, number of actions, and self-reported effort. The findings reveal that disengagement is associated with lower test performance and that process data-based measures such as response time and number of actions are more reliable indicators of engagement than self-reported effort. Gender disparities in disengagement further highlight the need for targeted interventions to improve test-taking motivation across diverse student populations.

Beyond test-taking behavior, two studies focus on improving measurement methodology. Yin et al. (in this issue) introduce an LCA-based approach to setting cut scores for context scales addressing challenges posed by skewed response distributions. By applying their method to PIRLS 2021 data, they demonstrate its potential to enhance the interpretability of context scales and provide a more statistically robust alternative to conventional judgment-based cut-score definitions. Demirkaya et al. (in this issue) examine latent profiles of mathematical skills by comparing student classifications derived from achievement and ability assessments using widely administered state assessments in the United States. Their study reveals substantial differences in the profiles emerging from these two classification approaches, highlighting the importance of using multiple measures to identify students with distinct instructional needs. These findings have direct implications for gifted education, as they suggest that relying on a single measure may overlook students who demonstrate strong cognitive potential despite lower achievement scores.

The final study in this special issue, by Altiner Sert and Arikan (in this issue) explores the relationship between social-emotional learning (SEL) and mathematics achievement, using data from the OECD's 2019 Survey on Social and Emotional Skills. Their findings suggest that emotional regulation and open-mindedness positively predict math performance, while high social engagement is negatively associated with achievement. Notably, SEL skills have a stronger predictive impact on students from lower socioeconomic backgrounds, reinforcing the importance of SEL programs in mitigating educational inequities.

**Concluding Thoughts**

The studies featured in this special issue demonstrate the evolving landscape of large-scale assessments, driven by advancements in data science, psychometric techniques, and a deeper understanding of student behavior. The findings highlight the increasing role of process data in improving assessment validity and fairness, the need for refined measurement models that accommodate diverse student populations, and the growing recognition of non-cognitive factors in shaping academic performance.

Future research should continue to explore AI-driven models for personalizing test accommodations, enhancing test development process and refine process data methodologies to improve engagement detection and response behavior analysis, and further examine the role of non-cognitive skills in educational assessments. As educational systems continue to embrace computer-based assessment practices and AI-driven methodologies, the intersection of assessment technology, psychometrics, and behavioral insights will remain a critical area of research. This special issue aims to inspire further innovation and interdisciplinary collaboration, ultimately contributing to more equitable and insightful large-scale assessments.

**References**

Anghel, E., Wry, E., & von Davier, M. (2025). Affective, behavioral, and cognitive engagement in ePIRLS: a latent profile analysis. *Educational technology research and development*, 1-27.

Bezirhan, U., von Davier, M., & Grabovsky, I. (2021). Modeling item revisit behavior: Thehierarchical speed–accuracy–revisits model. *Educational and PsychologicalMeasurement, 81*(2), 363-387.

Binici, S., & Cuhadar, I. (2022). Validating performance standards via latent class analysis. *Journal of Educational Measurement, 59*(4), 502-516.

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners: the role of noncognitive factors in shaping school performance--a critical literature review*. Consortium on Chicago School Research. 1313 East 60th Street, Chicago, IL 60637.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608.

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In *Validation of score meaning for the next generation of assessments* (pp. 11-24). Routledge.

Organisation for Economic Co-operation and Development. (2021). Beyond academic learning: First results from the survey of social and emotional skills. *OECD Publishing.*

Teig, N. (2024). Uncovering student strategies for solving scientific inquiry tasks: Insights from student process data in PISA. *Research in Science Education, 54*(2), 205-224.

Templin, J., & Jiao, H. (2012). Applying model-based approaches to identify performance categories. In *Setting performance standards* (pp. 379-397). Routledge.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non response. *British Journal of Mathematical and Statistical Psychology, 73*, 83-112.

Williams, G. A., & Kibowski, F. (2016). Latent Class Analysis and Latent Profile Analysis. In *Handbook of methodological approaches to community-based research: qualitative, quantitative, and mixed methods* (pp. 143-151). Oxford University Press.

Wise, S. L. (2017). Rapid-guessing behaviour: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*, 52–61. https://doi.org/10.1111/emip.12165

.

This page is intentionally left blank.

www.iejee.com

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.

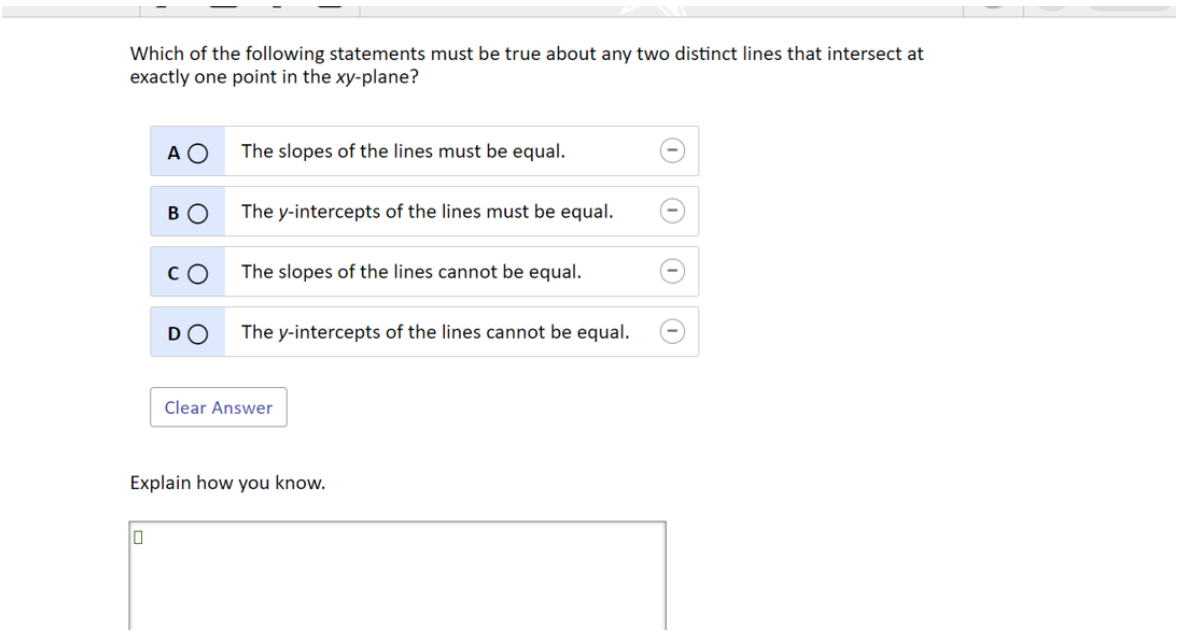
Sample Correct Responses

- 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.*

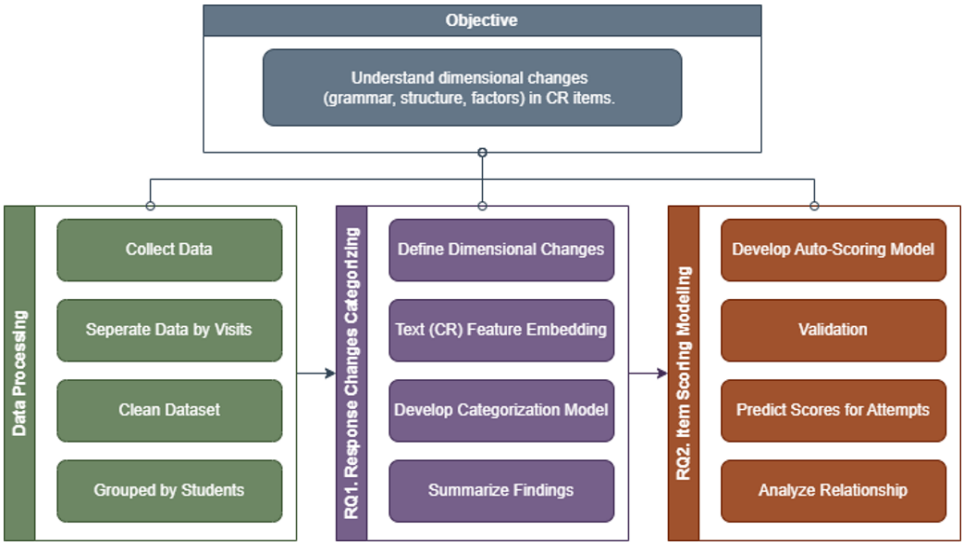
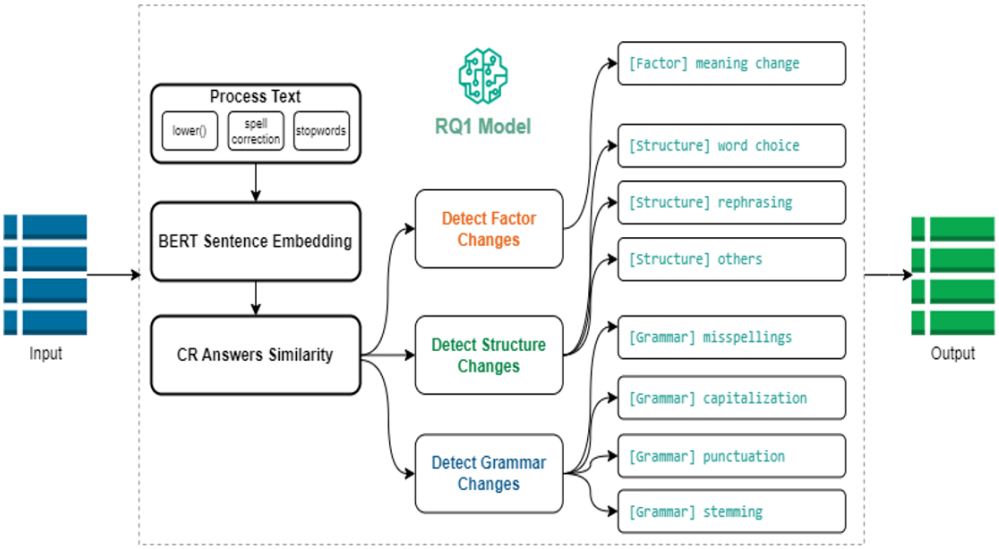



**Figure 3.**
*Model of the process developed for RQ1.*




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.*

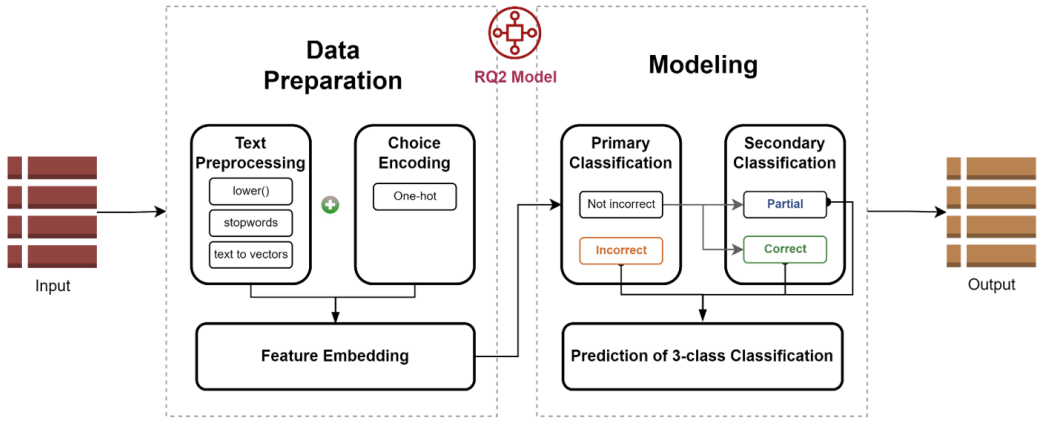
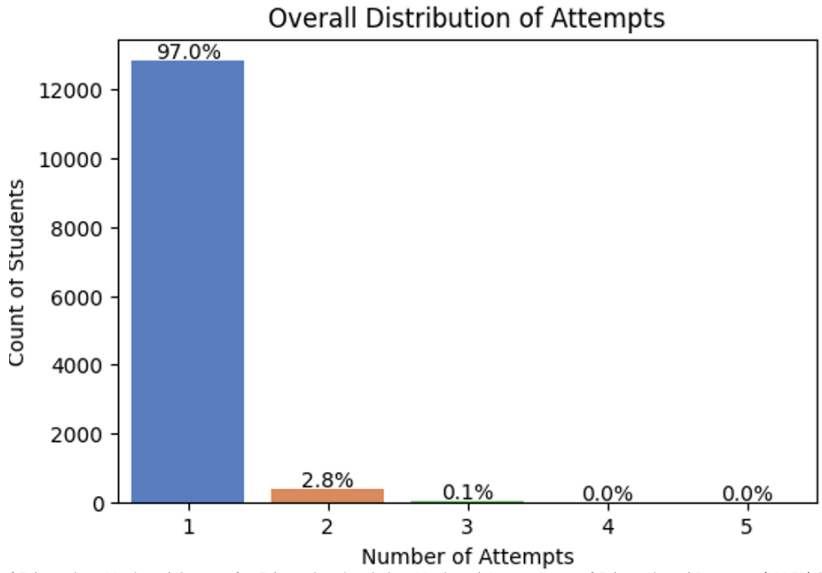



**Figure 5.**
*Number of students and their attempts to the selected item.*




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.*

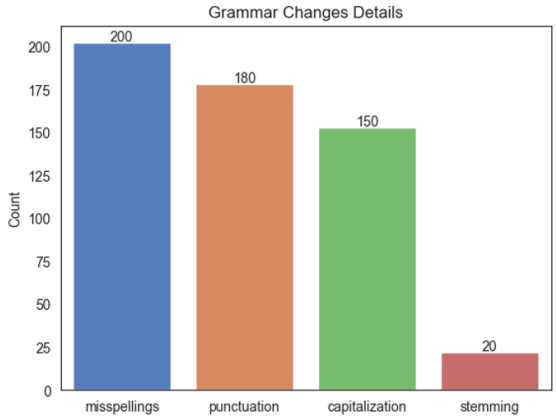
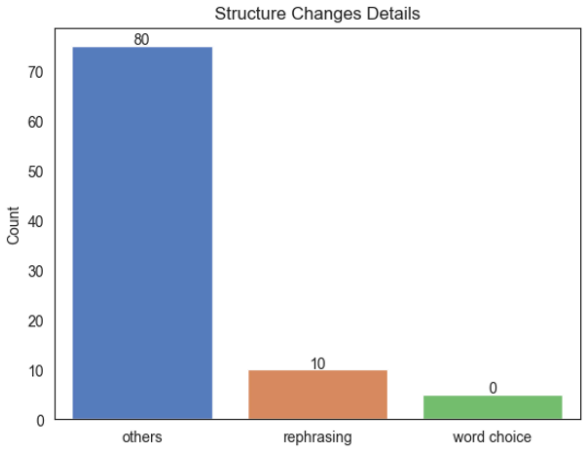



**Figure 6b.**
*Number of students with various types of structure changes.*




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).*

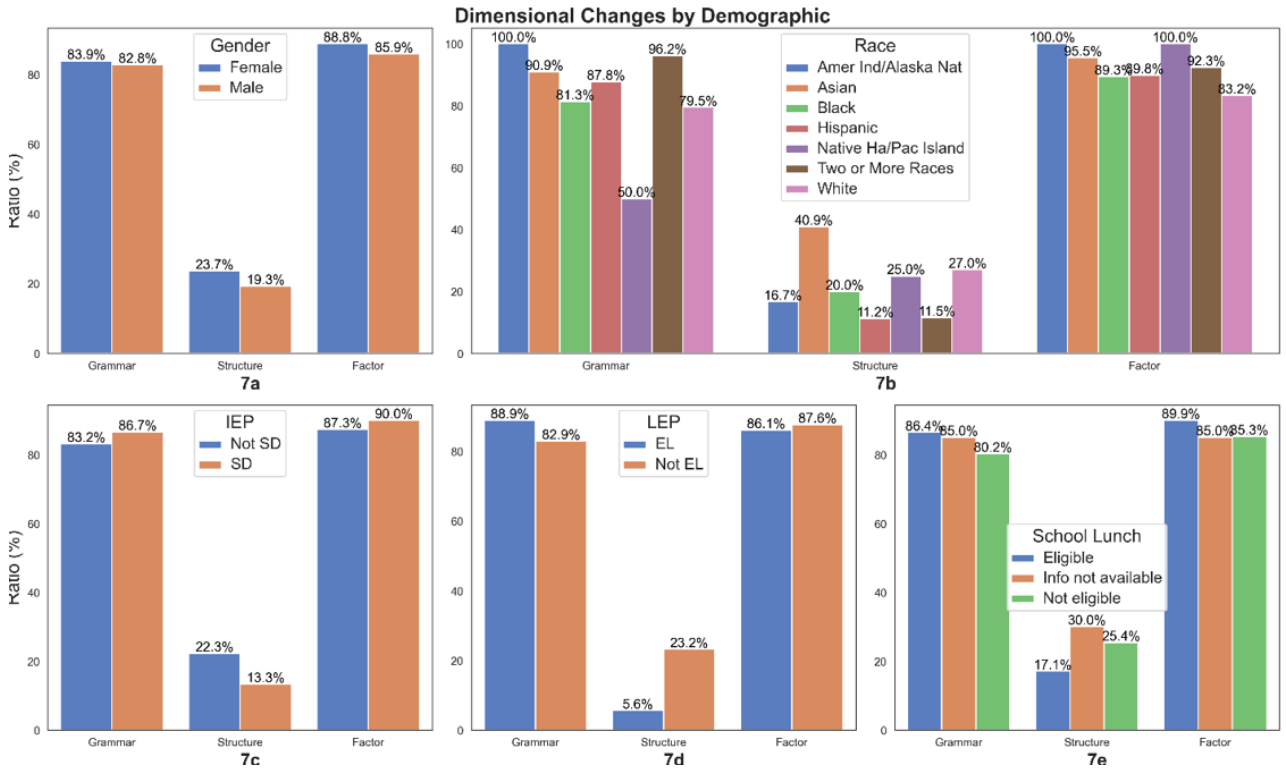


**Table 3.**
*Confusion matrix for the item scoring models performance.*

| | | 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*

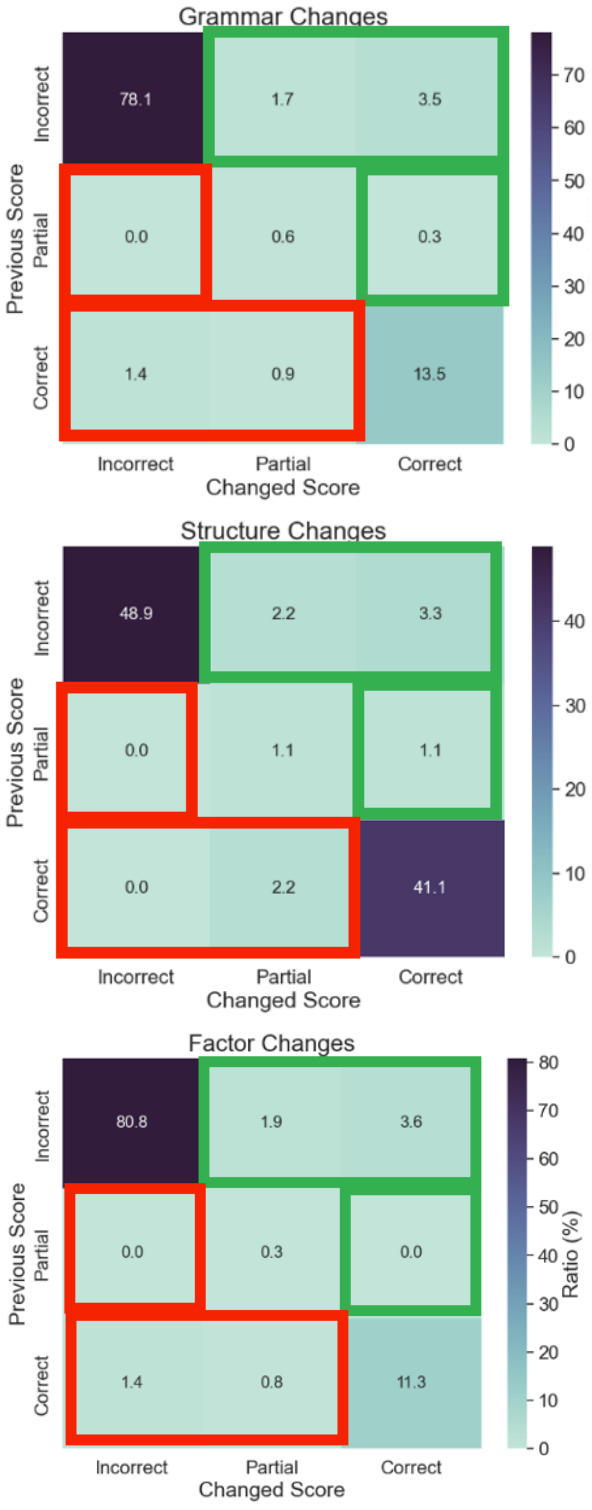
After training and evaluating, the item scoring model was applied to all attempts to generate predicted scores. These results provided insights into how students' scores changed between attempts. Given the concise nature of the written responses, it was anticipated that changes in the meaning of responses (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.

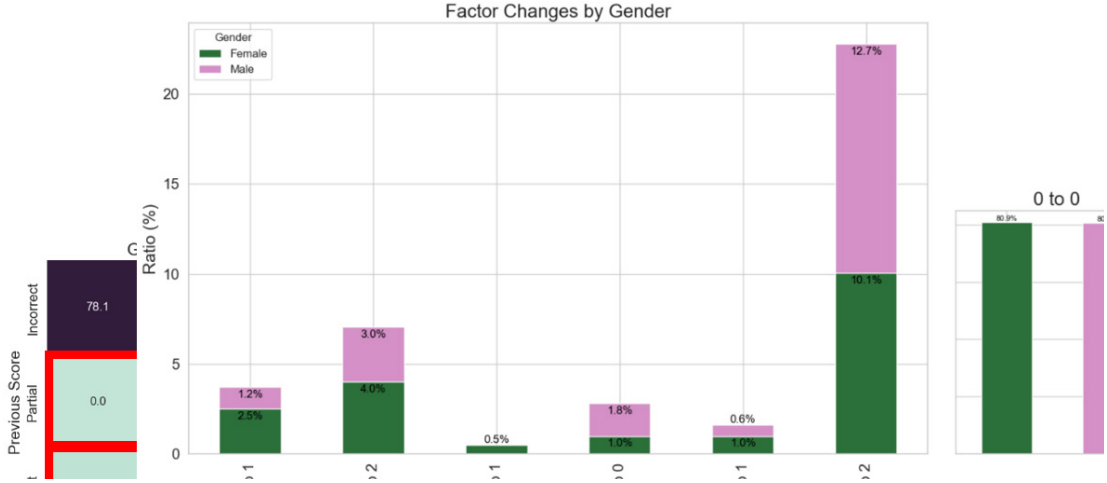
Analyzing factor changes by race reveals distinct patterns of score transitions among different demographic groups (Figure 10). Native Hawaiian/ Pacific Islander students exhibit the highest rate of improvement from "Incorrect" to "Partial" (0 to 1) at 25.0%. For the transition from "Incorrect" to "Correct" (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.**

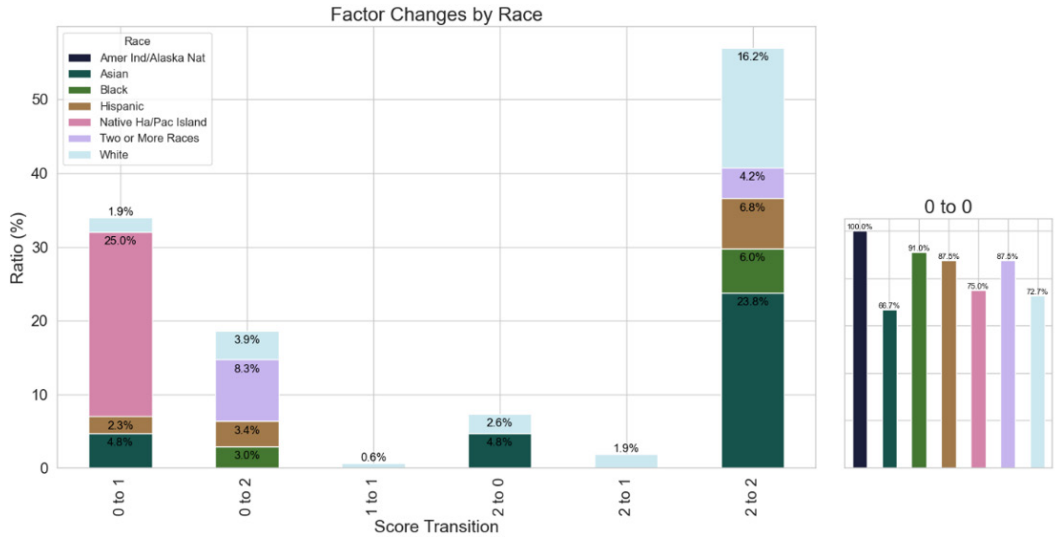
*Score transition patterns in factor dimension by gender.*




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*.

iejee

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

# Running Out of Time: Leveraging Process Data to Identify Students Who May Benefit from Extended Time

Burhan Ogut[a,*], Ruhan Circi[b], Huade Huo[c], Juanita Hicks[d], Michelle Yin[e]

Received : 12 September 2024
Revised : 27 December 2024
Accepted : 2 March 2025
DOI : 10.26822/iejee.2025.376

[a,*] **Corresponding Author:** Burhan Ogut, American Institutes for Research, Arlington, VA, USA.
E-mail: : bogut@air.org
ORCID: https://orcid.org/0000-0003-1729-1396

[b] Ruhan Circi, American Institutes for Research, Arlington, VA, USA.
E-mail: rcirci@air.org
ORCID: https://orcid.org/0000-0003-3854-1796

[c] Huade Huo, American Institutes for Research, Arlington, VA, USA.
E-mail: hhuo@air.org
ORCID: https://orcid.org/0009-0004-5014-646X

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

[e] Michelle Yin, Northwestern University, Illinois, USA.
E-mail: michelle.yin@northwestern.edu
ORCID: https://orcid.org/0000-0001-9333-1535



## Abstract

This study explored the effectiveness of extended time (ET) accommodations in the 2017 NAEP Grade 8 Mathematics assessment to enhance educational equity. Analyzing NAEP process data through an XGBoost model, we examined if early interactions with assessment items could predict students' likelihood of requiring ET by identifying those who received a timeout message. The findings revealed that 72% of students with disabilities (SWDs) granted ET did not use it fully, while about 24% of students lacking ET were still actively engaged when timed out, indicating a considerable unmet need for ET. The model demonstrated high accuracy and recall in predicting the necessity for ET based on early test behaviors, with minimal influence from background variables such as eligibility for free lunch, English Language Learner (ELL) status, and disability status. These results underscore the potential of utilizing early assessment behaviors as reliable predictors for ET needs, advocating for the integration of predictive models into digital testing systems. Such an approach could enable real-time analysis and adjustments, thereby promoting a fairer assessment process where all students have the opportunity to fully demonstrate their knowledge.

**Keywords:**

Extended Time Accommodation, NAEP Assessment, Process Data, Machine Learning, Test-Taking Behavior, Equitable Accommodations.

## Introduction

During the 2021-22 academic year, approximately 7.3 million students—or 15% of all public-school students in the United States—received special education services under the Individuals with Disabilities Education Act (IDEA), marking an increase from 13% in 2010-11 (De Brey et al., 2023). This growing demographic underscores the critical need to refine educational assessments to ensure they accurately reflect the abilities of students with disabilities. Most educational assessments are administered under standardized conditions, including the content, scoring, and administration, to guarantee that the results reflect students' abilities and not differences in assessment conditions.

Although standardized assessments aim to ensure fairness, they may inadvertently compromise the validity of test scores for students with disabilities (SWDs) by introducing construct-irrelevant variance—elements of the assessment process that are unrelated to the skills or knowledge being tested. Accommodations such as extended time (ET), sign language interpreters, and braille are implemented to mitigate construct-irrelevant variance by tailoring the administration format to the unique needs of SWDs, thereby facilitating a more equitable assessment environment (Bolt & Thurlow, 2006).

Federal law mandates the provision of accommodations for students with disabilities on both federal and statewide assessments to promote fairness and validity. However, despite legal requirements, the implementation and decision-making process regarding these accommodations often lacks clear, empirically-based guidelines. Individualized Education Program (IEP) teams, which include parents, regular education teachers, and special education teachers, have the responsibility to determine appropriate accommodations for each student with disabilities but often do so without sufficient data or guidance on their effectiveness or appropriateness (Hollenbeck, 2005).

Extended time has been shown to significantly improve the performance of students with disabilities, such as those with learning disabilities, ADHD, or anxiety disorders by allowing them to better demonstrate their knowledge and skills without the pressure of time constraints (e.g., Elliott & Marquart, 2004; Lovett, 2010). Potential mechanisms for the influence of extended time on students' performance include reduction in test-related stress, increased confidence and motivation (Alster, 1997; Elliott & Marquart,2004; Lovett & Leja, 2013),

When students who need extra time to complete an assessment are not provided with this accommodation, their performance may suffer significantly. Under time pressure, these students might start getting anxious and lose confidence and motivation. They may also rush to answer questions, a phenomenon known as speededness (Lu & Sireci, 2007). All these issues challenge the validity of the assessment results.

Although theoretically possible, removing all time constraints from assessments is impractical. Instead, we argue that monitoring students' progress during an assessment to identify those falling behind can allow for timely interventions. The timing of such interventions is crucial; too early, and it risks misidentifying students who do not require extra time, while too late can mean students have already hastened their responses to their detriment. This study seeks to find a balanced approach to when and how to grant additional time based on model fit statistics, thus determining the ideal point during an assessment to make these critical decisions (Lipnevich & Panaderom, 2021).

The introduction of digitally-based assessments opens new possibilities for more precisely tracking and analyzing students' test-taking behaviors through process data. This data can provide valuable insights into how accommodations are used and the extent to which they are effective. By employing advanced machine learning techniques to analyze process data from digital assessments, this study aims to not only enhance our understanding of how students utilize ET but also refine the decision-making process regarding its allocation. This innovative approach has the potential to make educational assessments more adaptive and inclusive, ensuring that they truly reflect student competencies and support equitable educational outcomes, fully aligning with the federal mandate for accessibility and fairness in educational testing.

*Relevant Literature*

The existing body of literature on ET accommodations reveals complex interactions between accommodations and test performance across various domains, including mathematics, reading, and college entrance exams. The review of the literature by Sireci et al. (2005) gave support to the interaction hypothesis, positing that while SWDs benefit from ET, students without disabilities (SWODs) do not. A differential boost in test performance favoring SWDs has also been documented (Fuchs et al., 2005; Gregg & Nelson, 2012), indicating that ET can significantly impact the fairness and equity of testing outcomes.

Despite these findings, traditional studies have predominantly relied on paper-based assessments, which do not provide granular data on how test-takers interact with test items and the testing environment. The introduction of digitally based assessments has begun to shift this landscape. The use of process data from digital platforms allows for a nuanced analysis of test-taker behaviors, including time management and problem-solving strategies (Lee & Haberman, 2015; van der Linden, 2019). This digital transition is critical as it provides an empirical basis for examining the temporal dimensions of test-taking, such as differential speediness (van der Linden et al., 1999) and the use of accessibility supports (Lee et al., 2021).

Notably, previous research has shown that SWDs often exhibit slower response times in both cognitive and academic tasks compared to their non-disabled peers, highlighting the relevance of ET (Wolff et al., 1990; Ofiesh et al., 2005). However, response time effort (RTE) measures, which assess the effort and motivation behind responses (Wise & Kong, 2005),

have been underutilized in the context of accessibility and accommodation research, especially in digital settings.

One significant gap in the literature is the reliable and valid identification of students who would benefit most from ET accommodations. Lovett (2010) critiqued the existing methods for determining eligibility for ET accommodations, which often rely on subjective judgments or diagnostic labels, pointing to a need for more objective and data-driven approaches. This research presents meaningful advancements to the existing literature on ET and digital educational assessments. By employing advanced machine learning techniques to analyze NAEP process data, this study aims to uncover patterns of ET use during assessments and addresses a crucial gap by offering an empirical, data-driven methodology for assessing the applicability of ET accommodations. This contributes significantly to the digital transformation of our education systems and the pursuit of equitable educational practices.

*Current Study*

This study had three primary objectives to enhance our understanding of ET usage in digital assessments through process data analysis. Firstly, we sought to provide empirical evidence supporting the use of ET accommodations by analyzing the typical extent of usage and profiling the characteristics of students who avail themselves of ET. Secondly, we investigated whether there are discernible differences in test-taking behaviors—such as task interaction, time allocation on individual items, and accommodation usage—among students when engaged with the assessment. Lastly, we employed predictive analytics to identify students at risk of not completing the assessment within the designated time, while they were still in the early stages of the assessment. The study was driven by the following research questions:

1. How is ET accommodation utilized by students, and does this usage vary according to the type of disability?

2. Are there observable differences between students with and without ET accommodations in interacting with the assessment (e.g., time spent on tasks and the number of actions performed)?

3. Can initial task engagement behaviors, such as time spent on tasks and student actions, predict which students may require ET accommodations?

**Methods**

*Data*

In this study, we analyzed two restricted-use datasets from the 2017 NAEP Grade 8 Mathematics assessment: process data and response data. The National Assessment of Educational Progress (NAEP) is the foremost national assessment, providing a comprehensive and ongoing evaluation of the knowledge and skills of students from both public and private schools throughout the United States across various academic subjects. With the transition to digital assessments in 2017, NAEP began collecting new types of data, allowing for detailed insights into student behavior during assessments. This includes metrics such as the duration students spend on tasks, their problem-solving approaches, and the utilization of available tools or features (National Center for Education Statistics, 2023). The process data for this analysis included records from an assessment block comprising approximately 28,000 participants. The NAEP response data encompasses information from the student background questionnaire, responses to cognitive items (i.e., mathematics assessment questions), teacher surveys, and school surveys. After processing and cleaning the process data, it was merged with the response data using student-level unique identifiers (i.e., pseudo IDs). Approximately 2 percent of the records were excluded from the analysis due to data quality issues, such as interrupted assessment sessions.

*Measures*

In the National Assessment of Educational Progress (NAEP), students granted the ET accommodation are allowed up to three times the standard time allocated for the assessment block. For the Grade 8 mathematics assessment, this translates to 90 minutes for students with ET accommodations, compared to the standard 30 minutes for those without. To identify students who, while not eligible for ET accommodations, might benefit from additional time, we focused on those unable to complete the assessment within the allotted period. We employed two primary measures for this analysis: one based on response data (i.e., ET accommodation status) and another on process data (i.e., ET accommodation usage).

*Process Data Measures:*

Extended Time Usage: We categorized students who were granted ET accommodations into those who utilized ET and those who did not, based on their total assessment time. Students exceeding the 30-minute limit (1800 seconds) were considered to have used ET.

Timeout Message: During the digital assessments conducted on tablets or laptops, a "timeout message" alerts students that their time has expired. This feature is critical for identifying students who might benefit from ET despite not being eligible. We analyzed the occurrence of timeout messages received by students while actively engaged in a task, using process data to determine whether the student was actively working

at the time of expiration. A binary indicator was then created to identify these students as potentially needing ET.

Measures of Student Interaction with the Assessment. We recorded the time and action related measures of students' interaction with the assessment for each math assessment item they attempted. Since NAEP allows students to navigate through the assessment in any order, including skipping items, we could not rely on item order as they appeared in the assessment for these measures. Instead, we defined "interaction" as referring to student entering and exiting any item. If a student revisits the same item again, under this definition, we recorded that interaction as separate from the earlier interaction with the same item. Therefore, in our analyses the item interaction order does not correspond to item order as they appear in the assessment. Using "interaction" variable that is agnostic to item order enabled us to control for students' preferences in interacting with the assessment items.

Early Interactions: We focus on the first 10 items, as analyzing these initial interactions offers an optimal balance between the timing of the additional time appraisal and the accuracy in identifying students likely to exhaust their allotted time.Exit Time and Actions: For the first 10 item interactions, we defined "exit time" as the total time a student spent from the start to the end of the current item interaction. We also tracked "actions" taken during each interaction, such as modifying a response or adjusting text in open-ended questions. The total number of actions, encompassing selecting options, focusing or defocusing on text fields, calculator key presses, and scratch work adjustments, was calculated for each item interaction to gauge student engagement levels.

Frequently Accessed Items: We identified items that were most frequently accessed by students during specific interactions, providing insights into item preferences and engagement patterns.

*Response Data Variables*

Not Reached Items: The concept of a "not reached" item, which stems from traditional paper-and-pencil assessments, is used by NAEP to identify items that a student did not respond to due to time constraints. Unlike the process used in paper assessments, NAEP does not utilize process data to determine not reached items. Instead, it assesses the responses at the end of an item block; if a student has one or more missing responses to subsequent items, those items are classified as "not reached."

Item Type: Information regarding the item type, such as multiple-choice single select or match multiple select, is extracted from the response data. This helps in understanding how different item types might affect the time needed and the strategies used by students during the assessment.

Demographics: Detailed demographic data, including disability status, English language learner status, eligibility for free or reduced-price lunch, specific types of disability, and whether ET was provided as an accommodation, are gathered from the response data. This information is crucial for both descriptive analyses, which aim to outline the characteristics of the study population, and predictive analyses, which seek to identify factors influencing the need for accommodations like ET.

**Analysis**

We utilized descriptive statistics and predictive analytics to address the research questions posed in this study. Initially, we extracted timing and interaction data from the process data. Using descriptive statistics, we conducted t-tests to explore patterns of ET usage and students' interactions with the assessment, focusing on both the general student population and specifically on students with disabilities. We also analyzed the relationship between the number of interactions with an item, the average cumulative time spent before exiting the item, and the number of actions taken by students.

For investigating the predictors of ET usage, we implemented machine learning-based predictive analytics. The dependent variable in these analyses was a binary indicator representing whether a student was actively interacting with an item when the time expired. The independent variables included demographic data such as English Language Learner (ELL) status, Disability status, the provision of ET accommodations, eligibility for free or reduced-price lunch, and various measures derived from process data that depicted students' interactions with the assessment.

Our predictive modeling began with logistic regression as a baseline approach. To enhance the robustness of our findings, we also utilized the XGBoost model, a decision-tree-based ensemble technique employing a gradient-boosting framework, noted for its effectiveness in various studies (Chen & Guestrin, 2016; Sahin, 2020; Osman et al., 2021). We tested multiple models incorporating different sets of timing and action variables to identify students who were more likely to benefit from ET accommodations by predicting those at risk of receiving a timeout message during the assessment. The models' hyperparameters were meticulously optimized using Bayesian Optimization (Nogueira, 2014) to enhance predictive accuracy, as detailed in Table 1 of our results section.

**Table 1.**

*XGBoost Hyperparameters Used in the Analysis*

| Hyperparameter | Bounds Used |
|---|---|
| Step size shrinkage used in update to prevents overfitting (learning_rate). | [0.01, 0.3] |
| Number of gradient boosted trees. Equivalent to number of boosting rounds (n_estimators). | [50, 500] |
| The maximum depth of a tree (max_depth). | [3, 10] |
| Control the balance of positive and negative weights, useful for unbalanced classes (scale_pos_weight). | [1, 5] |

Note: Hyperparameter names are in parentheses. Additional details on XGBoost hyperparameters can be found at https://xgboost.readthedocs.io/en/stable/parameter.html.

Bayesian Optimization requires a target score to evaluate the model's predictive power. Expanding upon the concept of the F-measure, which is calculated as the harmonic mean of precision and recall, we utilized the Fbeta-measure. The Fbeta-measure, or $F_\beta$, includes a configurable parameter known as beta.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

In our analysis, we adopted a larger beta value (beta=2), which inherently emphasizes recall over precision in our evaluation metrics. Specifically, this adjustment places less emphasis on precision—the proportion of students who were actually engaged with an item at the time of timeout among those identified—and more on recall—the proportion of correctly identified students who were engaged at timeout among all such students. This approach, denoted as the F2 score, aims to maximize the identification of students who could benefit from ET accommodations.

We partitioned the analytical dataset into two subsets, utilizing 80% of the data for training and reserving 20% as test data. The features for the predictive models included the exit times from the first 10 task interactions, the number of actions within the first 10 minutes, and various student demographic factors (such as whether ET was granted, eligibility for free or reduced-price lunch, special education status, and English Language Learner status).

To optimize the model's parameters, we conducted a 5-fold cross-validation combined with Bayesian Optimization on the training data. After determining the best hyperparameters, we applied both logistic regression and XGBoost models to the training dataset and evaluated their performance on the test dataset, which helped assess the models' generalizability beyond the training data. For interpretation of the machine learning models, we utilized SHapley Additive exPlanations (SHAP) values, which provide insights into the contribution of each feature to the predictive outcomes (Lundberg et al., 2020).

## Results

In the composition of our analytical sample, approximately 10% of the participants were SWDs, with more than half of these students identified as having specific learning disabilities, as detailed in Table 2. Other prevalent disabilities within the sample included speech impairments, emotional disturbances, and autism. In the subsequent sections, we present and discuss the findings corresponding to each of our research questions.

### Extended Time Usage (RQ 1)

As indicated in Table 2, among all students who were granted ET accommodations, only 25.1% utilized it. SWDs exhibited a slightly higher usage rate of ET at 27.4%, compared to 25% among SWODs. Usage rates among SWDs varied, ranging from 22.2% for students with intellectual disabilities (ID) to 28.1% for students with specific learning disabilities (SLD); however, these differences were not statistically significant.

Regarding the time spent on the assessment block, students, on average, spent 1462.10 seconds (approximately 24.37 minutes). Those without ET accommodations spent an average of 1444.45 seconds (around 24.07 minutes), while those with ET accommodations spent significantly more time, averaging 1681.44 seconds (about 28.02 minutes). Detailed minimum and maximum times spent are available in Table S1 in the supplemental files.

Subgroup analysis revealed variations in time spent on the assessment across different student categories. Among SWODs, those with ET accommodations took notably longer—1807.30 seconds (approximately 30.12 minutes)—compared to their peers without accommodations, who took 1446.70 seconds (about 24.11 minutes). SWDs with ET accommodations spent an average of 1647.98 seconds (approximately 27.47 minutes), while those without accommodations used about 1395.26 seconds (around 23.25 minutes). Specifically, students with autism, emotional disturbance (ED), specific learning disabilities (SLD), and speech impairment (SI) all spent more time on the test when granted ET accommodations compared to those without. The most significant difference was observed in students with SI, where those with ET used 1798.53 seconds (approximately 29.98 minutes) versus 1420.57 seconds (about 23.68 minutes) for those without.

Further, we examined the prevalence of timeout messages and items marked as "not reached" during the assessment, comparing across disability types and the use of ET accommodations. Table 3 illustrates that among all students, 23.62% of those without ET accommodations received a timeout message, a stark contrast to only 1.41% of those with ET accommodations.

Similarly, 21.87% of students without accommodations did not reach one or more test questions, compared to 7.95% of those with ET accommodations. Among SWODs, 23.52% received timeout messages without ET accommodations, significantly reduced to 1.39% for those with accommodations. The pattern was similar for "not reached" items, with 21.75% of students without ET accommodations and 8.78% with ET accommodations failing to reach certain tasks.

SWDs showed a similar trend, with 25.87% without ET accommodations receiving timeout messages, compared to only 1.41% of those with accommodations. For "not reached" items, 24.59% of SWDs without ET accommodations did not reach tasks, significantly reduced to 7.73% among those with ET accommodations. When analyzed by specific disability types, all groups—including those with autism, ED, hearing impairment (HI), intellectual disability (ID), SLD, and SI—demonstrated lower rates of timeout messages and not reaching tasks when provided with ET accommodations. For instance, autistic students without ET accommodations had 23.53% receiving timeout messages and 22.06% not reaching certain tasks, which dramatically decreased to 0% and 10%, respectively, with ET accommodations. These patterns of reduction were consistent across the other disability types, underscoring the significant benefits of ET accommodations in reducing timeouts and instances of incomplete tasks, thus enabling a more thorough assessment of student knowledge and capabilities.

*Assessment Interactions of students with ET and without ET accommodation (RQ 2)*

Table 4 offers an in-depth overview of the most frequently accessed items during the assessment, detailing the item type, average exit time, and the number of actions during the first ten interactions with any item. It also highlights variations based on whether students received a timeout message. Given the flexibility of the assessment format, students can interact with items in a non-linear order, potentially revisiting earlier items to revise their responses after gaining clearer insights from subsequent questions.

The initial interaction typically involved VH356842, a non-cognitive item focusing on completion directions. Students without a timeout message completed this task in an average of 10.88 seconds (approximately 0.18 minutes) with 3.18 actions, while those who received a timeout message took slightly longer, exiting at an average of 12.37 seconds (about 0.21 minutes) with a comparable number of actions (3.23).

During the second interaction, the most engaged item was VH266695, a multiple-choice single select (MCSS) item. Students without a timeout message spent an average of 46.01 seconds (about 0.77 minutes) with 6.12 actions. In contrast, those with a timeout message took longer, exiting the task after an average of 62.01 seconds (approximately 1.03 minutes) and performing more actions (7.90).

The third interaction frequently involved VH304549, a match multiple select (MatchMS) item. Students without a timeout message exited this task in 102.64 seconds (roughly 1.71 minutes) with 11.00 actions, whereas those with a timeout message took longer, exiting at an average of 132.24 seconds (about 2.20 minutes) with 11.91 actions.

This pattern was consistent across all interactions, with students receiving timeout messages consistently exiting items later and engaging in more actions than those without such messages. By the tenth interaction, involving another MatchMS item, VH261992, students without a timeout message averaged an exit time of 579.19 seconds (about 9.65 minutes) with 11.88 actions. Conversely, those who received a timeout message took significantly longer, exiting at an average of 820.26 seconds (approximately 13.67 minutes) and taking 15.54 actions. These findings indicate that students who spend more time and interact more extensively with tasks are more likely to encounter timeout messages.

*Identifying Students who may Need ET (RQ 3)*

The results from the logistic regression models, which predicted the probability of encountering a timeout message based on students' interactions with tasks, are detailed in the supplemental file (Table S2). Generally, the logistic regression models exhibited lower accuracy compared to the XGBoost models (Figure 1). Consequently, we selected the XGBoost model for further analysis.

The findings from the XGBoost analysis (Table 5) highlighted the complex balance between the timeliness of detecting a student who will receive a timeout message and the accuracy of this detection, demonstrating high accuracy, high recall rate, and a significant F2 score. This table presents the results of 10 models, each employing a distinct subset of interaction-specific variables, refined through manual recursive feature addition. While all models consistently incorporate background variables, the first model focuses exclusively on data from the interaction with the first item and does not integrate subsequent information from later items. In contrast, the model analyzing the interaction with the tenth item includes all background data and information from all previous interactions.

Although it is feasible to develop additional models incorporating variables from interactions beyond the first 10 items, focusing on these initial interactions provides an optimal balance between the timing of the additional time appraisal and the accuracy of identifying students likely to exhaust their allotted time.

**Table 2.**

*Time Spent on Math Assessment Block (in Seconds) by Disability Type and Use of Extended Time Accommodation*

| Student's Identified Disability Type | Percent of sample | Percent of Using ET | Average Time Spent | | |
|---|---|---|---|---|---|
| | | | All students | Students without ET Accommodation | Students with ET Accommodation |
| All Students | 100 | 25.10% (0.27) | 1462.10 (2.42) | 1444.45 (2.15) | 1681.44* (17.72) |
| Students without Disabilities | 90.15 | 25.00% (0.28) | 1452.95 (2.27) | 1446.70 (2.18) | 1807.3* (40.11) |
| Students with Disabilities | 9.83 | 27.4%† (1.35) | 1546.45 (12.94) | 1395.26 (12.09) | 1647.98* (19.65) |
| Autism | 0.53 | 25.00% (5.29) | 1523.49 (53.35) | 1389.66 (45.4) | 1637.24* (89.19) |
| Emotional Disturbance | 0.68 | 23.8% (4.68) | 1419.94 (43.11) | 1283.46 (49.82) | 1530.17* (64.96) |
| Hearing Impairment | 0.14 | 25.00% (11.2) | 1535.70 (89.06) | 1370.30 (104.48) | 1656.00 (129.73) |
| Intellectual Disability | 0.27 | 22.20% (8.15) | 1314.46 (59.83) | 1336.56 (72.66) | 1301.77 (84.97) |
| Specific Learning | 5.41 | 28.10% (1.85) | 1530.28 (16.56) | 1418.18 (16.04) | 1603.24* (24.97) |
| Speech Impairment | 0.79 | 26.10% (4.19) | 1606.08 (45.39) | 1420.57 (35.61) | 1798.53* (80.89) |

Notes: Standard errors in parentheses. Developmental delay, orthopedic impairment, brain injury, visual impairment, other health" issues or "other write-in" disabilities were excluded from this table. Percent using ET is calculated for those who used it more than 30 mins.
* Statistically significant difference (<.05) compared to students without ET accommodations.
† Statistically significant difference compared to SWODs.
Data source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

**Table 3.**

*Percent of Students (SE) Receiving Timeout Messages or Having "Not Reached" Items by Disability Type and Use of Extended Time Accommodation*

| Student's Identified Disability Type | Overall | | Students without ET Accommodation | | Students with ET Accommodation | |
|---|---|---|---|---|---|---|
| | Timeout Message | Not Reached | Timeout Message | Not Reached | Timeout Message | Not Reached |
| All Students | 21.97 (0.25) | 20.83 (0.24) | 23.62 (0.27) | 21.87 (0.26) | 1.41* (0.26) | 7.95* (0.6) |
| SWODs | 23.14 (0.27) | 21.52 (0.26) | 23.52 (0.27) | 21.75 (0.56) | 1.39* (1.32) | 8.78* (0.29) |
| SWDs | 11.24 (0.61) | 14.51 (0.68) | 25.87 (0.26) | 24.59 (1.36) | 1.41* (1.3) | 7.73* (0.66) |
| Autism | 10.81 (2.56) | 15.54 (2.99) | 23.53 (5.18) | 22.06 (5.07) | - | 10.00 (3.38) |
| Emotional Disturbance | 13.16 (2.3) | 15.79 (2.71) | 23.81 (4.68) | 27.38 (4.89) | 0.96* (0.96) | 7.69* (2.63) |
| Hearing Impairment | 8.11 (5.56) | 10.81 (5.99) | 25.00 (11.2) | 31.25 (12) | 4.55* (4.55) | 4.55* (4.55) |
| Intellectual Disability | 11.27 (3.19) | 13.41 (3.63) | 22.22 (8.15) | 14.81 (6.97) | - | 8.51 (4.11) |
| Specific Learning | 21.97 (0.82) | 20.83 (0.88) | 26.4 (1.81) | 23.35 (1.74) | 1.43* (0.39) | 6.94* (0.84) |
| Speech Impairment | 23.14 (2.34) | 21.52 (2.72) | 26.13 (4.19) | 32.43 (4.46) | 0.93* (0.94) | 7.48* (2.55) |

Notes: Standard errors in parentheses. Developmental delay, orthopedic impairment, brain injury, visual impairment, other health" issues or "other write-in" disabilities were excluded from this table.
-Suppressed due to small sample size.
* Statistically significant difference (<.05) compared to students without ET accommodations.
Data source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

**Table 4.**

*Most Frequently Accessed Task and Task Type, Average Exit Time (in Seconds), and Number of actions during the First 10 Interactions by Receipt of Timeout Message*

| Interaction Number | Most Frequently Interacted task | Task Type | Exit Time | Number of all actions | Exit Time without Timeout Message | Number of all actions without Timeout Message | Exit Time with Timeout Message | Number of all actions with Timeout Message |
|---|---|---|---|---|---|---|---|---|
| 1 | VH356842 | Directions† | 11.20 (0.12) | 3.19 (0.01) | 10.88 (0.15) | 3.18 (0.02) | 12.37 (0.16) | 3.23 (0.02) |
| 2 | VH266695 | MCSS | 49.52 (0.26) | 6.51 (0.06) | 46.01 (0.28) | 6.12 (0.07) | 62.01 (0.58) | 7.90 (0.16) |
| 3 | VH304549 | MatchMS | 109.10 (0.37) | 11.20 (0.06) | 102.60 (0.4) | 11.00 (0.06) | 132.24 (0.90) | 11.91 (0.13) |
| 4 | VH336968 | FillInBlank | 184.70 (0.55) | 22.34 (0.17) | 174.10 (0.57) | 21.53 (0.17) | 222.32 (1.40) | 25.19 (0.44) |
| 5 | VH303873 | MatchMS | 248.50 (0.72) | 7.782 (0.07) | 232.60 (0.72) | 7.37 (0.07) | 305.07 (1.87) | 9.25 (0.18) |
| 6 | VH263651 | GridMS | 330.60 (0.92) | 13.92 (0.15) | 307.10 (0.9) | 12.88 (0.16) | 414.31 (2.38) | 17.61 (0.4) |
| 7 | VH304553 | MatchMS | 416.00 (1.1) | 10.98 (0.06) | 385.70 (1.08) | 10.56 (0.06) | 523.56 (2.85) | 12.46 (0.2) |
| 8 | VH262355 | FillInBlank | 500.80 (1.29) | 19.76 (0.19) | 461.40 (1.24) | 18.55 (0.20) | 640.35 (3.29) | 24.05 (0.51) |
| 9 | VH287980 | MCSS | 562.80 (1.38) | 8.164 (0.08) | 516.90 (1.32) | 7.49 (0.07) | 725.98 (3.47) | 10.55 (0.27) |
| 10 | VH261992 | MatchMS | 632.20 (1.5) | 12.68 (0.12) | 579.20 (1.42) | 11.88 (0.13) | 820.26 (3.70) | 15.54 (0.33) |

Notes: † This is non-cognitive task ptoviding the directions for the assessment. Standard errors in parentheses. "MCSS" stands for "Multiple Choice Single Select" item. "MatchMS" stands for "Match Multiple Select" item. "FillInBlank" stands for "Fill in the Blank" Item. "GridMS" stands for "Grid Multiple Select" item. Data source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

**Table 5.**

*Analysis of True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), F2 Score, Accuracy, and Recall by Interaction Number in the XGBoost Model*

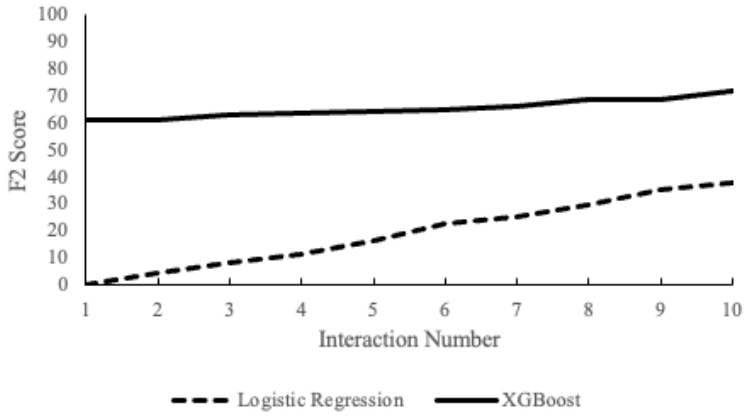
| Interaction number # | TP | TN | FP | FN | F2 Score | Accuracy | Recall |
|---|---|---|---|---|---|---|---|
| 1 | 1230 | 530 | 3750 | 30 | 61.48 | 31.79 | 98.01 |
| 2 | 1020 | 2060 | 2220 | 240 | 61.47 | 55.58 | 80.88 |
| 3 | 1050 | 2050 | 2240 | 200 | 63.36 | 55.97 | 83.90 |
| 4 | 1040 | 2180 | 2100 | 210 | 63.93 | 58.27 | 83.19 |
| 5 | 1010 | 2470 | 1810 | 250 | 64.34 | 62.85 | 80.40 |
| 6 | 990 | 2670 | 1620 | 270 | 64.80 | 66.01 | 78.73 |
| 7 | 1020 | 2690 | 1600 | 240 | 66.48 | 66.81 | 80.88 |
| 8 | 1030 | 2850 | 1440 | 230 | 68.47 | 69.86 | 81.67 |
| 9 | 1000 | 3070 | 1220 | 260 | 68.95 | 73.36 | 79.52 |
| 10 | 1040 | 3080 | 1210 | 210 | 71.69 | 74.40 | 83.03 |

Note: Using 20% of the sample as the testing set. All sample sizes are rounded to the nearest 10.
Data source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

The metrics used to evaluate the models included true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), F2 score, accuracy, and recall, assessed across various interaction numbers. TP refers to students correctly identified by the model as having received a timeout message, while TN indicates students who did not receive a timeout message and were correctly identified as such. FP represents students incorrectly predicted to receive a timeout message, and FN refers to students who did receive a timeout message but were mistakenly predicted not to have received one. These metrics allow for a comprehensive evaluation of the model's effectiveness in classifying students based on their timeout status.

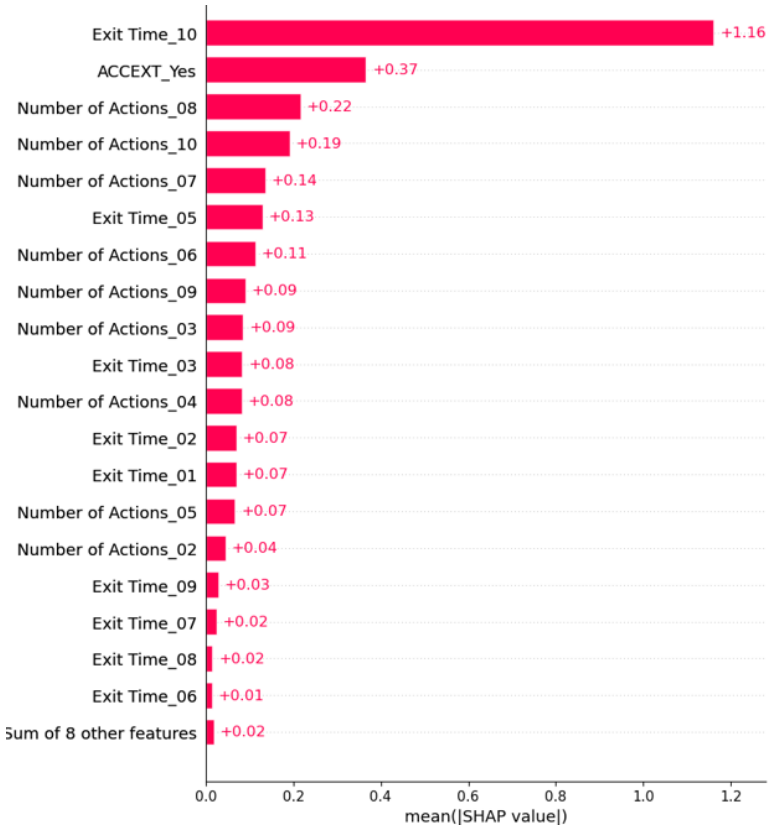
During the interaction with the first item, the model demonstrated a high recall rate of 98.01%, successfully identifying 1,230 TPs. It achieved an accuracy of 31.79% and an F2 score of 61.48, indicating a strong ability to

**Figure 1.**
*Prediction Accuracy (F2 Scores) for Logistic Regression and XGBoost models by Interaction Number*




Data source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

**Figure 2.**
*The Mean Absolute SHAP Value for All Features*




Data source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

identify students who received a timeout message. However, this came at the cost of a high number of false positives, with 3,750 students incorrectly classified as receiving a timeout message. By the interaction with the second item, the model's accuracy had improved to 55.58%, the recall rate adjusted to 80.88%, and the F2 score remained stable at 61.47, showcasing the model's evolving efficiency in more accurately predicting timeout incidents as more interaction data became available.

The model's performance continued to improve through the interactions with the third to tenth items. By the third task, accuracy had slightly increased to 55.97%, recall rose to 83.90%, and the F2 score reached 63.36. With the fourth task, there was a notable improvement in accuracy to 58.27%, although the recall rate slightly decreased to 83.19%, with the F2 score climbing to 63.93.

As the model processed data from the fifth through seventh items, accuracy consistently improved,

peaking at 66.81% by the seventh task. The recall rate remained stable around 80%, with the F2 score progressively increasing to 66.48. The subsequent interactions, from the eighth to the tenth items, further underscored the model's enhanced accuracy, which reached 74.40% by the tenth task. After a brief dip in recall to 81.67% on the eighth item, it rebounded to 83.03% by the tenth, accompanied by an increase in F2 scores to 71.69.

This progression highlighted the delicate balance between early detection and maintaining high recall and F2 scores. Early detection, pivotal in identifying students likely to receive a timeout message in initial interactions, improved as more interaction data was integrated, thereby enhancing overall model accuracy while sustaining a commendable recall rate and F2 score. This demonstrated the XGBoost model's capacity to effectively identify students who would benefit from ET accommodations early in the assessment process.

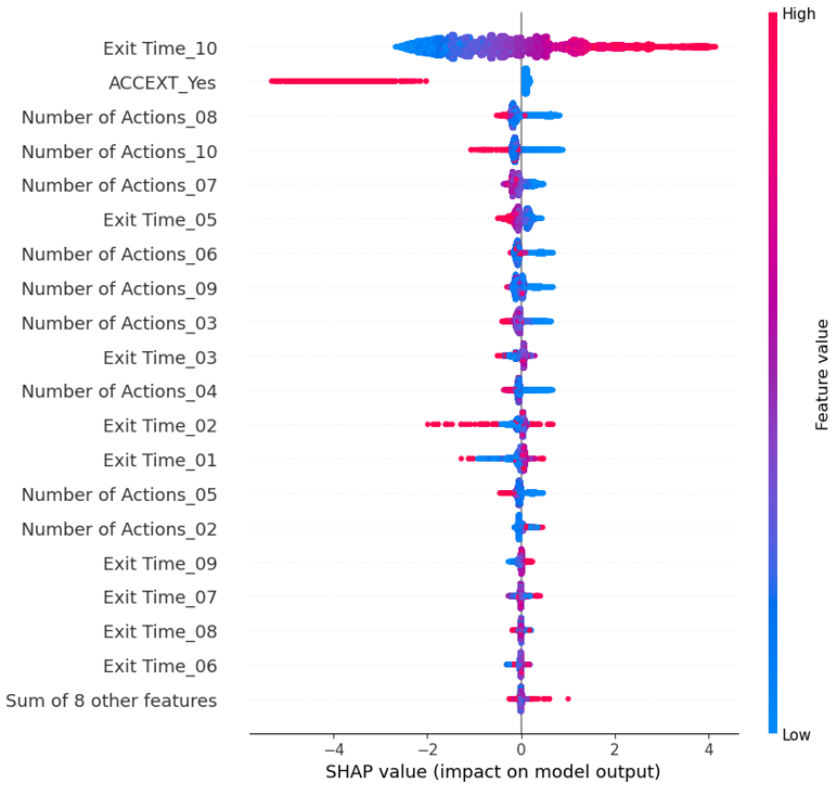
The SHAP (Shapley Additive exPlanations) values, a game-theoretic approach designed to explain the output of machine learning models (Lundberg, et al., 2020), were used in interpreting the influence of model features on predictions. For the 10th model, we examined the SHAP values through various visualizations. Figure 2 displayed the mean absolute value of the SHAP values for each predictor, emphasizing the importance of the time of exit for the interaction with the 10th item, availability of ET accommodations, and the number of all actions recorded during the 8th item as key influences on the model's predictions.

Each dot in the Beeswarm plot (Figure 3) represents an individual student, with the horizontal position indicating the impact magnitude of each feature on the model's predictive accuracy for that student. This visualization aids in understanding how different features influence the likelihood of a timeout message. For example, students with ET accommodations (represented in red) were less likely to receive a timeout message compared to those without ET accommodations (in blue). The plot also illustrates the distribution of effect sizes, notably the long right tails for the "exit time on the interaction with the 10th task" feature, indicating significant variability in how this particular variable impacts the model's predictions.

In exploring individual cases, Figures 4, 5, 6, and 7 illustrate the contribution of each feature to the model's output, shifting it from the base value—representing the average output over the training dataset—to specific outcomes for true positives (TP, Figure 4), true negatives (TN, Figure 5), false positives (FP, Figure 6), and false negatives (FN, Figure 7). Features that increase the likelihood of a specific prediction are shown in red, while those that decrease the likelihood

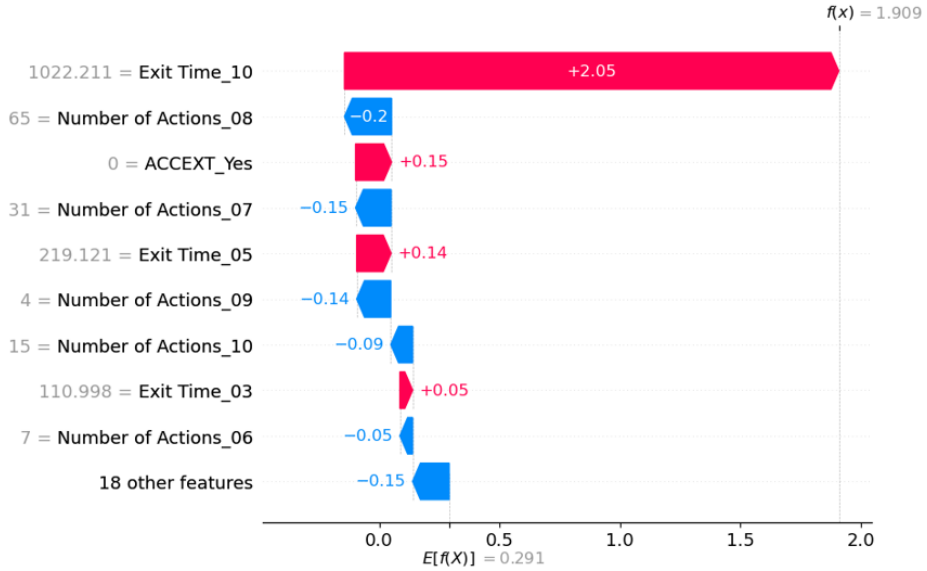
**Figure 3.**

*Beeswarm Plot Showing How Exit Time, Extended Time Accommodation, and the Number of Actions Drive Model's Prediction*



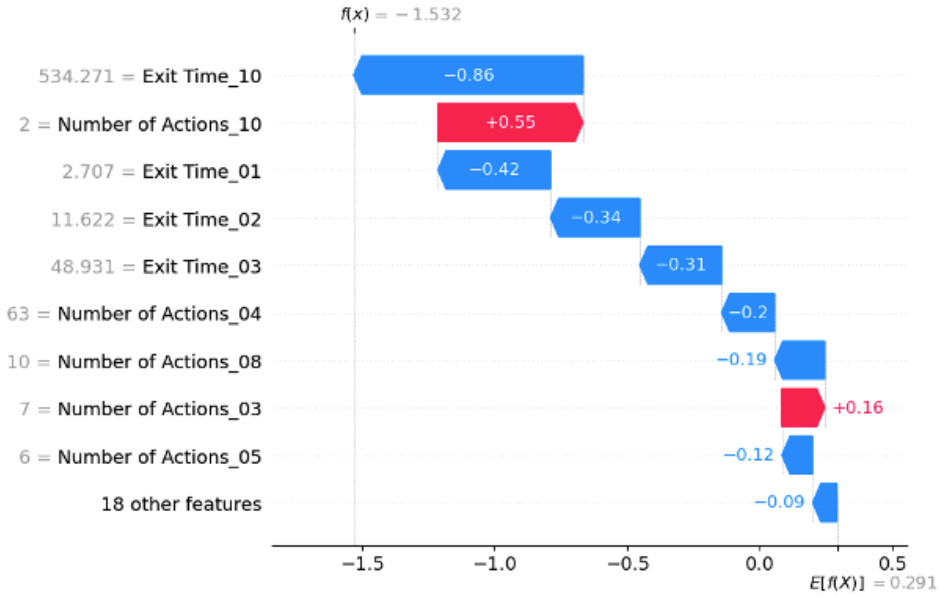

Data source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

**Figure 4.**
*Waterfall Plot Demonstrating How Individual Features Contribute Towards True Positives (TP)*




Note. The red bars represent features that push the prediction higher, such as the exit time for the interaction with the 10th task.
Data Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

**Figure 5.**
*Waterfall Plot Demonstrating the Contribution of Individual Features Towards True Negatives (TN)*

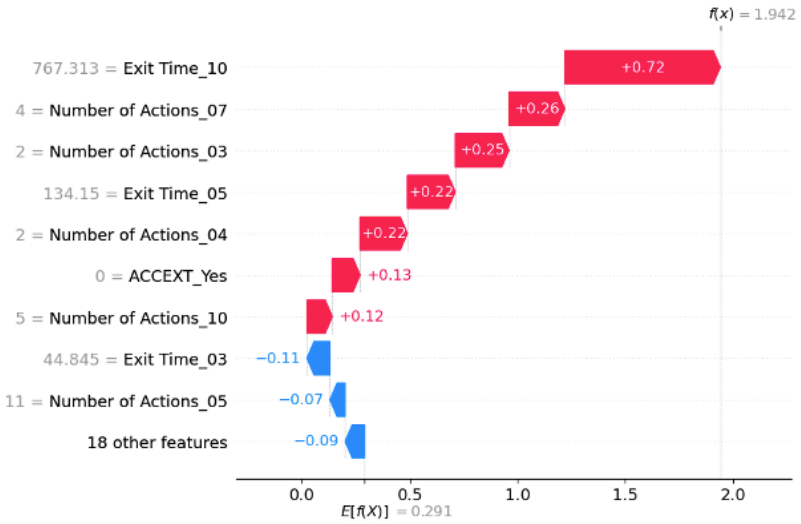



Note. Blue bars represent features that lower the prediction, such as the exit time for the interaction with the 10th task.
Data source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

are depicted in blue. For example, a longer exit time during the interaction with the 10th item, specifically 1022.21 seconds (approximately 17.04 minutes), is highlighted in Figure 4. This feature significantly elevates the probability of a student being classified as having received a timeout message, reflecting its positive influence on the prediction (depicted in red). Conversely, a shorter exit time for the same item, recorded at 534.27 seconds (about 8.9 minutes) as shown in Figure 5, significantly reduces the likelihood of being classified as receiving a timeout message, shown in blue.

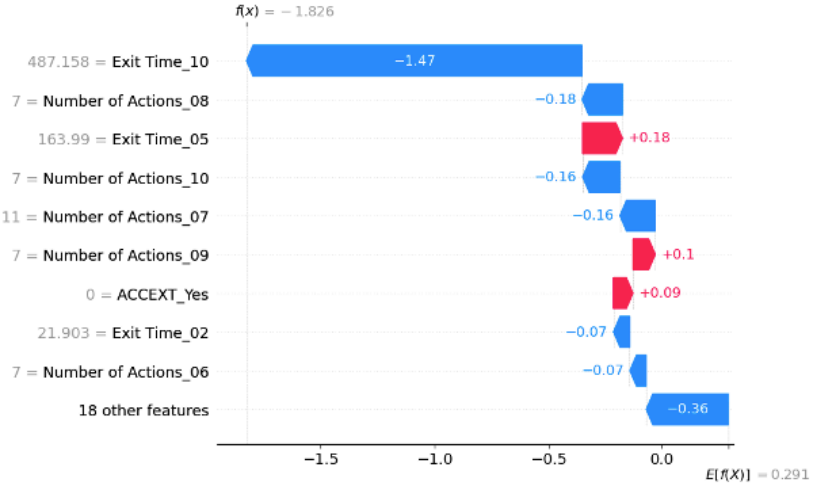
Notably, the exit time for the 10th item also plays a critical role in the misclassification of cases, influencing both false positives and false negatives. This is evident in Figures 5 and 6, where the impact of shorter or longer exit times, respectively, steers the model's predictions, affecting its accuracy in identifying true versus false outcomes. These visualizations underscore the importance of this particular feature in shaping the model's predictions and highlight the potential for refining predictive accuracy by further analyzing the implications of interaction times and other influential variables.

**Figure 6.**
*Waterfall Plot Demonstrating the Role of Individual Features Towards False Positives (FP)*




Note. Figure highlights how certain features like the exit time for the interaction with the 10th task can also lead to misclassification.
Data source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

**Figure 7.**
*Waterfall Plot Demonstrating How Specific Features Contribute Towards False Negatives (FN).*




Note. Figure shows the significant influence of the exit time for the interaction with the 10th task on misclassifications.
Data source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

## Discussion

This study investigated the utilization of ET accommodations among SWDs using process data from the 2017 NAEP Grade 8 Mathematics assessment. We explored the potential of early assessment interactions as predictors for the necessity of ET accommodations.

Extended time is a commonly granted accommodation (Frey & Gillispie, 2018); however, our findings indicate that, in the context of large-scale assessments, only about 12 seconds beyond the allotted 30 minutes were used by those granted ET. Remarkably, approximately 72% of SWDs granted ET did not utilize it at all, with usage varying from under a minute to nearly an hour among those who did. On the other hand, about 24% of students without ET were actively engaged with tasks when they received a timeout message, highlighting a significant unmet need for ET among the tested population.

The variability in ET allocation across states, IEP teams, and schools of differing socioeconomic statuses (Lovell, 2020) underscores the challenges in the current approach to granting accommodations. These disparities, coupled with our findings of unused ET and instances of students working on assessment when time expired, point to the need for a more objective and timely method of identifying students who truly need ET. The timing of this identification is crucial; it should be early enough to prevent increased anxiety, lower motivation and rushed test-taking but also accurate in pinpointing those in need. Our results

suggest that student behavior in the initial minutes of an assessment is a viable early indicator of ET necessity. Employing the XGBoost model, we achieved high accuracy and recall in identifying these students, highlighting the model's practical application in early identification.

Furthermore, our analysis identified specific factors that significantly influence the need for ET. Notably, the exit time during the 10th item interaction, the availability of ET accommodations, and the number of actions during the 8th item interaction were strong predictors. Interestingly, students' background variables such as eligibility for free lunch, ELL status, and disability status had minimal impact on the model's predictive power, promoting educational equity by not overemphasizing demographic factors.

Our study contributes to the literature on the use of process data and predictive analytics in educational assessments, supporting the development of adaptive testing designs and the analysis of differential test-taking speeds among diverse student groups (van der Linden, 2019; Lee & Chen, 2011). The ability to predict ET needs based on early test behavior marks a significant step toward more equitable testing practices. Nearly a quarter of students without ET accommodations could benefit from them, suggesting profound implications for their academic success.

The implications of our findings are important for educational policy and practice, particularly for the NAEP assessments, which biennially evaluate student performance nationwide. The most recent NAEP mathematics assessment, administered in 2022, includes a wide demographic with approximately 116,200 grade 4 students and 111,000 grade 8 students. The findings suggest that educators and testing organizations need to reevaluate the provision of extended-time accommodation. A predictive approach based on early assessment behavior can help identify students who might otherwise be missed, thus ensuring that all students can demonstrate their knowledge fully and equitably. This proactive approach can help shape future guidelines on ET accommodations, fostering a more inclusive digital education environment.

Additionally, our study demonstrated the effectiveness of machine learning models, specifically the XGBoost model, in handling complex educational data. These models could be incorporated into digital testing systems to provide real-time analysis and predictions about students' needs for accommodations, further improving the fairness of these assessments.

Future research should expand this methodology to other subjects and grade levels to broaden understanding of ET accommodations across various educational contexts. Additionally, investigating the impact of receiving a timeout message on first block of a NAEP assessment on performance in the second block of NAEP assessment and integrating students' performance in the early-stages of the assessment with process data variables could provide deeper insights into pacing strategies and the overall assessment experience. This study represents an initial effort to guide further exploration in educational assessment, aiming to foster more inclusive and equitable testing environments.

**Acknowledgments**

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324P210002 to the American Institutes for Research (AIR). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

**References**

Elliott, S. N., & Marquart, A. M. (2004). Extended Time as a Testing Accommodation: Its Effects and Perceived Consequences. Exceptional Children, 70(3), 349-367. https://doi.org/10.1177/001440290407000306

Alster, E. H. (1997). The Effects of Extended Time on Algebra Test Scores for College Students With and Without Learning Disabilities. Journal of Learning Disabilities, 30(2), 222-227. https://doi.org/10.1177/002221949703000210

Lovett B. J., Leja A. (2013). Students' perceptions of testing accommodations: What we know, what we need to know, and why it matters. Journal of Applied School Psychology, 29, 72–89.

Bolt, S. E., & Thurlow, M. L. (2006). *Item-level effects of the read-aloud accommodation for students with reading disabilities (Synthesis Report 65).* National Center on Educational Outcomes, University of Minnesota. Retrieved from https://files.eric.ed.gov/fulltext/ED495897.pdf.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *22nd acm sigkdd international conference on knowledge discovery and data mining*, (pp. 785-794).

De Brey, C., Zhang, A., & Dillow, S. (2023). *Digest of Education Statistics 2021 (NCES 2023-009).* Washington, DC.: National Center for Education Statistics.

Fuchs, L. S., Fuchs, D., & Capizzi, A. M. (2005). Identifying appropriate test accommodations for students with learning disabilities. *Focus on Exceptional Children*, *37*(6).

Gregg, N., & Nelson, J. M. (2012). Meta-analysis on the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities*, *45*(2), 128–138. https://doi.org/10.1177/0022219409355484

Lee, D., Buzick, H., Sireci, S. G., Lee, M., & Laitusis, C. (2021). Embedded Accommodation and Accessibility Support Usage on a Computer-Based Statewide Achievement Test. *Practical Assessment, Research & Evaluation*, 26, 25.

Lee, Y. H., & Haberman, S. J. (2016). Investigating test-taking behaviors using timing and process data. *International Journal of Testing*, 16(3), 240-267.

Lipnevich, A & Panadero, E. (2021). A review of feedback models and theories: descriptions, definitions, and conclusions. *Frontiers in Education, 6* (2021), 10.3389/feduc.2021.720195

Lovett, B. J. (2020). Disability Identification and Educational Accommodations: Lessons From the 2019 Admissions Scandal. *Educational Researcher*, *49*(2), 125-129. https://doi.org/10.3102/0013189X20902100

Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research*, 80(4), 611–638.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, 26(4), 29–37.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., . Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 1, 2522-5839.

National Center for Education Statistics (2023, September). *Process Data From the 2017 NAEP Grade 4 and Grade 8 Mathematics Assessment*. https://www.nationsreportcard.gov/process_data/

Nogueira, F. (2014). *Bayesian Optimization: Open source constrained global optimization tool for Python*. Retrieved from https://github.com/fmfn/BayesianOptimization

Ofiesh, N., Mather, N., & Russell, A. (2005). Using speeded cognitive, reading, and academic measures to determine the need for extended test time among university students with learning disabilities. *Journal of Psychoeducational Assessment*, *23*(1), 35–52. https://doi.org/10.1177/073428290502300103

Osman, A. ,E., A., Ahmed, A.,N., Chow, M.,F., Huang,Y.,F., & El-Shafie, A. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor, Malaysia. *Ain Shams Engineering Journal*, 12*(2)*, 1545-1556. https://doi.org/10.1016/j.asej.2020.11.011.

Provasnik, S. (2021). Process data, the new frontier for assessment development: rich new soil or a quixotic quest?. *Large-scale Assessments in Education*, *9*(1), 1-17.

Sahin, E., K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences*, 2:1308. https://doi.org/10.1007/s42452-020-3060-1

Sireci S.G., Banda E., Wells C.S. (2018) Promoting Valid Assessment of Students with Disabilities and English Learners. In: Elliott S., Kettler R., Beddow P., Kurz A. (eds) *Handbook of Accessible Instruction and Testing Practices*. Springer, Cham. https://doi.org/10.1007/978-3-319-71126-3_15

Sireci, S. G., Scarpati, S. E., & Li , S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, 75(4), 457–490.

Thurlow, M. L., Elliott, J. L., & Ysseldyke, J. E. (2003). *Testing students with disabilities: Practical strategies for complying with district and state requirements*. Thousand Oaks, CA: Corwin Press.

Stone E.A. & Cook L.L. (2018) Fair Testing and the Role of Accessibility. In: Elliott S., Kettler R., Beddow P., Kurz A. (Eds.) Handbook of Accessible Instruction and Testing Practices. Springer, Cham. https://doi.org/10.1007/978-3-319-71126-3_4

Wolf, M. K., Hanwook Y., Guzman-Orth, D., & Abedi, J. (2022). Investigating the Effects of Test Accommodations with Process Data for English Learners in a Mathematics Assessment, *Educational Assessment, 27*(1), 27-45, DOI: 10.1080/10627197.2021.1982693

iejee

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

# Investigating the Differential Relationship Between the Big Five Domains of Social and Emotional Skills and Mathematics Achievement

Mihriban Altiner Sert[a,*], Serkan Arıkan[b]

Received : 4 November 2024
Revised : 13 January 2025
Accepted : 5 March 2025
DOI : 10.26822/iejee.2025.377

[a,*] **Corresponding Author:** Mihriban Altiner Sert, Enka Schools, Istanbul, Türkiye.
E-mail: mihriban.altiner@gmail.com
ORCID: https://orcid.org/0000-0003-4186-8124

[b] Serkan Arıkan, Faculty of Education, Bogazici University, Istanbul, T,rkiye.
E-mail: serkan.arikan1@bogazici.edu.tr
ORCID: https://orcid.org/0000-0001-9610-5496

## Abstract

The current study explores the differential relationship between social and emotional learning (SEL), based on the Big Five personality traits, and mathematics achievement among Turkish high school students. Using data from the OECD's 2019 Survey on Social and Emotional Skills (SSES), it examines how SEL dimensions predict math outcomes and how these relationships vary by gender, socioeconomic status (SES), and level of SEL evaluation in schools. Key findings reveal that open-mindedness and emotional regulation positively correlate with math achievement, while high social engagement shows a negative association. Girls' SEL skills had a stronger predictive value for math achievement than boys, and SEL had a more substantial impact on students from lower SES backgrounds. Formal SEL assessment in schools was also related to higher math scores. These results emphasize the importance of SEL programs tailored to specific demographic needs, particularly for disadvantaged students, and suggest that formal SEL assessment in schools could enhance academic outcomes.

**Keywords:**

Social And Emotional Learning; Mathematics Achievement; Differential Relationship; Survey on Social And Emotional Skills

## Introduction

In recent years, the rapid advancement of digitalization and globalization has profoundly reshaped the educational landscape, necessitating a holistic approach to student development. Beyond the traditional focus on cognitive skills, there is a growing acknowledgment of the pivotal role that social and emotional learning (SEL) plays in equipping students with the competencies required to navigate a complex world (OECD, 2021). Social and emotional skills, defined as the consistent patterns of thoughts, feelings, and behaviors that individuals can cultivate through formal and informal learning experiences, are recognized as important determinants of socio-economic outcomes throughout one's life (OECD, 2021).

SEL is increasingly viewed as essential to both educational and social development. The Collaborative for Academic, Social, and Emotional Learning (CASEL) describes SEL as the process through which individuals, both young and old, acquire and apply the knowledge, skills, and attitudes necessary to develop healthy identities, manage emotions, achieve personal and collective goals, empathize with others, establish and maintain supportive relationships, and make responsible decisions (CASEL, 2020). In this digitalized era, having advanced social and emotional skills plays an important role in individuals' career development (Green, 2024). Individuals with advanced social and emotional skills, including assertiveness, creativity, and perseverance, are likely to have a more significant influence on the future labor market (OECD, 2024). Thus, the growing importance of SEL is evident in global educational initiatives aimed at fostering both cognitive and social-emotional competencies, thus enabling students to tackle the challenges of modern life.

*Importance of Social and Emotional Learning*

In accordance with the reforms introduced in education systems, particularly since the 2000s, it is evident that the relevance of social and emotional skills that contribute to enhancing academic performance has significantly increased (Candeias et al., 2020). Also, social and emotional skills provide a range of multidimensional advantages, facilitating students' growth across different aspects of their development including academic success, individual well-being, health and profession (Kankaraš & Suarez-Alvarez, 2019; OECD, 2024). Therefore, the role of SEL in education is increasingly recognized as critical for student success, both academically and in broader life outcomes. SEL fosters the development of essential life skills, including emotional regulation, empathy, decision-making, and relationship-building, which are indispensable for navigating the complexities of modern society (Goleman, 1995). Research has demonstrated that SEL programs can significantly enhance students' academic performance, social behaviors, and emotional well-being and it is highlighted that students who participate in SEL programs not only perform better academically but also exhibit more positive social behaviors and fewer behavioral issues (Durlak et al., 2011; Taylor et al., 2017). Also, it is reported that SEL programs provide an increase in life-satisfaction, more cooperative behavior, and more self-efficacy on students (Durlak et al., 2015; Gol-Guven, 2021).

To develop more targeted and focused educational interventions, it is essential to understand the relationship between SEL and math ability for various kinds of student groups, including gender groups, students from low- and high-socioeconomic-status (SES) backgrounds and student groups whose SEL is assessed to varying degrees. Research has consistently shown that gender plays a significant differential role in the development of social and emotional skills. For instance, girls often excel in social and emotional competencies such as empathy and cooperation, which are closely linked to academic success (Poropat, 2009). However, there is also evidence suggesting that boys may benefit differently from SEL programs, with some studies indicating that boys may show greater improvements in areas such as emotional regulation and task performance when exposed to targeted SEL interventions (Taylor et al., 2017). Exploring gender differences in the relationship between SEL and mathematics achievement has the potential to provide insights into how educational strategies can be tailored to support both boys and girls effectively.

SES is another critical factor influencing academic achievement. Students from higher SES backgrounds typically have access to more resources, both at home and in school, which can enhance their academic performance. Conversely, students from lower SES backgrounds often face additional challenges that can hinder their academic success, such as limited access to educational resources, lower parental involvement, and greater exposure to stress (Sirin, 2005). The OECD (2021) has emphasized the importance of addressing these disparities through targeted educational interventions that support the development of social and emotional skills, helping to level the playing field for students from disadvantaged backgrounds. Therefore, it is necessary to investigate how SES moderates the relationship between the Big Five domains of SEL and mathematics achievement, providing insights that could inform policies aimed at reducing educational inequalities.

Furthermore, the extent to which SEL is formally evaluated within schools significantly impacts its effectiveness. Schools that actively assess and promote SEL tend to foster environments that support both social and academic growth, whereas those that do not assess may miss critical opportunities to enhance student outcomes (CASEL, 2019). The significance of systematic SEL assessment in promoting academic achievement could be highlighted by analyzing the differential relationship of SEL based on whether students' social and emotional abilities are evaluated formally, informally, or not at all.

*Evaluating SEL as a Large-Scale Assessment*

Since assessment is a crucial component of comprehending a construct, SEL assessment is an important component of building social and emotional skills in order to create effective teaching techniques and learning outcomes (Agliati et al., 2020). A thorough SEL assessment system should be put in place, according to researchers, educators, and politicians from different countries, in order to support

student performance and achievement. Agliati et al. (2020) state that in order to support student learning in the classroom, social and emotional competences must be evaluated, just like the other learning domains. They contend that proper evaluation practices may give students feedback on their performance, assist them in monitoring their personal growth, and advise teachers on the best teaching methods to use.

In order to assess SEL globally, OECD is conducting a study on social and emotional skills. The most extensive international study on SEL skills is conducted by the OECD, which includes data from 10 cities across 9 nations. The objective was to create and provide a conceptual framework for the Social and Emotional Skills Study, which aims to clarify the educational, family-related, and personal elements that either facilitate or hinder the development of these abilities in a variety of student populations and environments (Kankaraš & Suarez-Alvarez, 2019). The current conceptual framework of Kankaraš and Suarez-Alvarez (2019) focuses on the underlying skills of the Big Five model that are indicative of positive life effects. It incorporates the merged and integrated competences from different applicable frameworks.

*Present Study*

The current study examines the differential relationship between the Big Five domains of social and emotional skills-task performance, emotional regulation, engaging with others, collaboration, and open-mindedness-and mathematics achievement. The study utilizes data from the OECD's 2019 Survey on Social and Emotional Skills (SSES) in Turkey, with a specific focus on 9th to 11th-grade students in Istanbul. The OECD's framework for social and emotional skills, which integrates various applied frameworks, underscores the critical role these skills play in shaping educational outcomes and overall life success (Kankaraš & Suarez-Alvarez, 2019).

In Turkey, the integration of SEL into the educational system has been less comprehensive compared to other countries, particularly concerning formal evaluation and curriculum integration. This study aims to bridge this gap by providing empirical evidence on how SEL, as conceptualized through the Big Five domains, is related to academic achievement in mathematics. By doing so, it offers valuable insights that could guide the development of more effective SEL programs in Turkish schools, potentially leading to improved educational outcomes for students across various demographics.

The current study is distinguished by its focus on the Turkish educational context, applying the Big Five model of social and emotional skills to predict mathematics achievement. While previous research has extensively explored the impact of SEL programs on general academic performance, there is a notable paucity of studies examining how these relationships may vary across different demographic and socioeconomic groups within Turkey. For instance, the meta-analysis conducted by Durlak et al. (2011) revealed that students participating in SEL programs exhibited enhanced academic performance, improved social behaviors, and reduced emotional distress. However, these studies have not sufficiently explored whether these benefits are consistent across gender, socioeconomic status (SES), or the level of SEL evaluation within schools. Therefore, the current study seeks to address these gaps by not only applying the Big Five domains to assess their relationship with mathematics achievement but also by investigating how this relationship may differ based on gender, SES, and the extent to which SEL is formally evaluated within schools. As the current study is one of the first studies conducted in Turkey in this area, it offers valuable insights that could inform educational policy and practice in the country. Additionally, the findings may contribute to the broader international discourse on the role of social and emotional skills in education, particularly in contexts where cultural and socioeconomic factors significantly influence educational outcomes.

The study is guided by the following research questions:

1. To what extent do the Big Five domains of SEL—task performance, emotional regulation, engaging with others, collaboration, and open-mindedness—predict mathematics achievement among high school students?

2. Is there a differential relationship between the SEL skills and mathematics achievement for gender groups and SES groups?

3. To what extent evaluation of social and emotional skills at school moderated the relationship between SEL and mathematics achievement?

**Method**

*Participants*

The current study used a dataset collected for the OECD's Survey on Social and Emotional Skills (SESS) study (OECD, 2021). The participants of the OECD's SESS study were 10- and 15-year-old students from 10 cities: Bogota, Colombia; Daegu, Korea; Helsinki, Finland; Houston, Texas, United States; Istanbul, Turkey; Manizales, Colombia; Moscow, Russian Federation; Ottawa, Ontario, Canada; Sintra, Portugal; and Suzhou, People's Republic of China. The OECD used a two-stage stratified random sampling method to choose the participants: first, schools were chosen, and then students were chosen from those schools.

In the current study 15-year-old students' data was used as older students could provide more consistent responses to self-assessment scales (Poropat, 2009; Rice and Pasupathi, 2010). Thus, the sample for this study comprised of 3168 students from 80 different high schools located in Istanbul.

*The Instruments*

The instruments and information listed below were used to produce the variables of the current study: a SEL survey, math achievement grades provided by schools, a survey assessing students' socioeconomic status, and a survey requesting information on how social and emotional skills were assessed in the classroom. The details are provided below.

*The survey for SES*

One of the instruments of the current study was the social-emotional skills survey of SSES 2019 study conducted by OECD. Based on the "Big Five Model" (John, Naumann, & Soto, 2008), the SSES theoretical framework was developed to assess the social and emotional competencies of youth (Chernyshenko et al., 2018). Collaboration, emotional regulation, engaging with others, open-mindedness and task performance, each with three subdimensions, were the five main domains of the SSES 2019 study.

In the survey the collaboration domain was defined as a combination of the abilities of empathy, trust, and cooperation that empathy is the ability of understanding and caring the other people and their well-beings; trust is the ability to assume that people generally act with good intentions and to forgive the wrong behaviors; co-operation is the ability to live together peacefully with others and respects the interdependence of all individuals (Kankaraš & Suarez-Alvarez, 2019). A sample item for the collaboration domain was "I am ready to help anybody."

Emotional regulation domain was defined as emotional stability with the combination of the skills; stress resilience, optimism, and emotional control that stress resilience is the ability to modulate anxiety effectively and solve problems calmly; optimism is the ability to have hopes for life positively and optimistically; emotional control is the ability to apply effective methods for controlling anger, aggression, and irritation in case of frustration (Kankaraš & Suarez-Alvarez, 2019). A sample item for the emotional regulation domain was "I keep my emotions under control".

Engaging with others domain was defined as extraversion with the combination of the skills; sociability, assertiveness, and energy that sociability is the ability to initiate and sustain social interactions with people; assertiveness is the ability to articulate thoughts, needs, and emotions with confidence and create social impact; energy is the ability to engage daily life with enthusiasm, energy, and spontaneity (Kankaraš & Suarez-Alvarez, 2019). A sample item for the engaging with others domain was "I like to spend my free time with others".

Open-mindedness domain was defined as openness to experience with the combination of the skills; curiosity, tolerance, and creativity that curiosity is the ability to have passion for learning, comprehension, and intellectual investigation; tolerance is the ability to be open to different perspectives and to appreciate the diverse values and cultures; creativity is the ability to generate innovative ways by means of vision, explorations, and learning from failure (Kankaraš & Suarez-Alvarez, 2019). A sample item for the open-mindedness domain was "I am willing to be friends with people from other cultures".

Task performance domain was defined as conscientiousness with the combination of the skills; responsibility, self-control, and persistence that responsibility is the ability to fulfill the commitments, as well as being punctual and trustworthy; self-control is the ability to resist disturbances and spontaneous desires and concentrate on the present task to reach a particular objective; persistence is the ability to persevere until a task or activity is completed (Kankaraš & Suarez-Alvarez, 2019). A sample item for the task performance domain was "I finish things despite difficulties in the way".

*Assessing mathematics achievement*

The current study used the standardized school grade for math classes taken in school as a proxy for mathematical achievement. Since participating cities have distinct grading systems, the OECD converted all grades to a scale of 1 to 50 (OECD, 2021).

*Socioeconomic status index*

The socioeconomic status (SES) index is derived from information about the household possessions (HOMEPOS), parental employment status as determined by the international socio-economic index of occupational status (ISEI), and parental education as determined by the International Standard Classification of Education scheme (ISCED). Open-ended questions were included in the surveys for parents and students to gather data on home possessions, occupation, and education. The authors of the current study divided the subjects into three equal-number groups to generate three categories based on socioeconomic status.

*Evaluation of social and emotional skills in schools*

Evaluations of social and emotional learning in the schools were another variable to take into account. By answering the following question, teachers disclosed information about whether social and emotional competencies were assessed in their institutions. "Is students' achievement in social and emotional skills evaluated in your school? No, we don't evaluate these skills; Yes, using informal evaluation (e.g. oral reports to students or parents, etc.); Yes, using formal evaluation (e.g. written reports, grades, etc.)" (OECD, 2021, p.12).

**Data Analysis**

First, the proposed model's fit was assessed in the current study using confirmatory factor analysis (CFA) (See Figure1). The main goal of confirmatory factor analysis is to statistically evaluate the significance of a hypothesized factor model, in other words, whether the sample data support hypothesized model (Schumacker & Lomax, 2004). Then, utilizing the main SEL domains, mathematical achievement was predicted using structural equation modeling (SEM). Lastly, SEM analyses were repeated for gender groups, SES groups and SEL evaluation groups to evaluate differential relationships. CFA and SEM analyses were performed using Mplus 7.4 to address the research questions of the current study. Sample weights can be taken into account by Mplus during the analysis process (Muthén & Muthén, 2012).

In CFA, 15 subdimensions of SEL were hypothesized to be related to five main domains (Collaboration, Emotional regulation, Engaging with others, Open mindedness and Task performance) as proposed in the Big Five Model. Goodness of fit indices show whether the data and the proposed model are similar. According to these goodness of fit indices, a good fit is indicated by a Root Mean Square Error of Approximation (RMSEA) value of 0.06 or less, and an acceptable fit is indicated by a value of 0.10 or less. A good fit is indicated by a Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) of 0.95 or higher, while an acceptable fit is indicated by a value of 0.90 or higher (Byrne, 1998; Ullman, 2001). After evaluating the fit of the measurement model, mathematical achievement was predicted by these five main domains of SEL using SEM. The analyses were repeated for the gender, SES, and SEL evaluation groups in order to assess whether the findings differ for various groups or not. MLR estimation method was used in CFA and SEM analysis as the achievement and SEL domains were created as continuous variables. The assumptions of normality, linearity and multicollinearity were evaluated, and it was concluded that none of the assumptions were violated.

**Results**

The major goal of the current study is to investigate the differential relationship between the social and emotional skills and mathematics achievement of students for various groups such as gender groups, SES groups and SEL evaluation groups. The following section contains the preliminary and comprehensive analyses for the research questions.

*Descriptive Statistics*

The descriptive statistics of mathematics achievement and social and emotional skill subdimensions were provided to indicate the key characteristics of the data (See Table 1). For the grouping variables, the frequencies are provided in Table 2.

*Confirmatory Factor Analysis of SEL Measurement Model*

The confirmatory factor analysis results for three competing models are presented in Table 3. In Model 1, 15 subdimensions were hypothesized to be related to

**Figure 1.**
*Measurement model of the study*

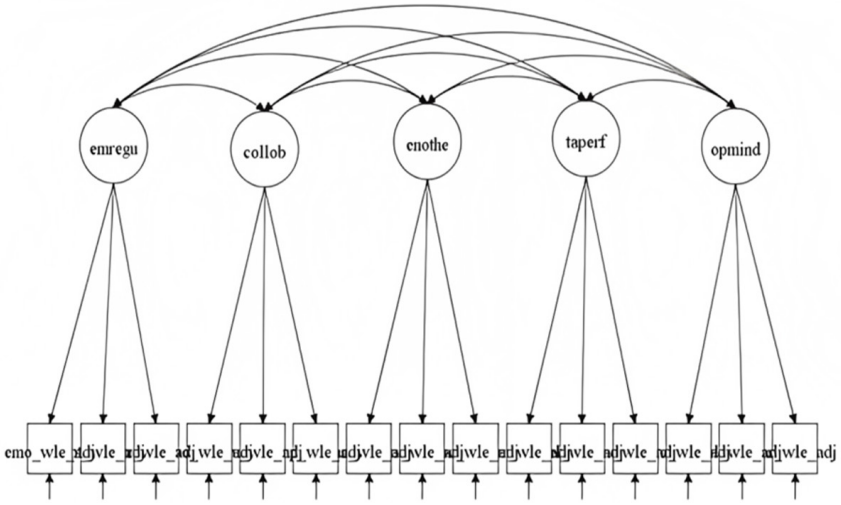


**Table 1.**
*Descriptive statistics of mathematics achievement and SEL subdimensions*

| | Mean | Median | Std. Deviation | Skewness | Kurtosis |
|---|---|---|---|---|---|
| Mathematics achievement | 29.15 | 28.86 | 10.52 | -.031 | -.864 |
| Collaboration by | | | | | |
| Empathy | 638.55 | 625.66 | 93.53 | .784 | 1.004 |
| Trust | 502.27 | 504.64 | 84.69 | -.376 | 2.231 |
| Cooperation | 627.54 | 617.99 | 85.54 | .610 | .608 |
| Emotional regulation by | | | | | |
| Emotional control | 512.30 | 510.40 | 88.72 | .135 | 3.516 |
| Optimism | 535.73 | 537.08 | 93.39 | .066 | 3.020 |
| Stress resilience | 512.09 | 514.62 | 111.85 | -.129 | 2.061 |
| Engaging with others by | | | | | |
| Assertiveness | 521.59 | 514.14 | 110.77 | .426 | .848 |
| Energy | 561.97 | 555.38 | 93.08 | .700 | 2.319 |
| Sociability | 583.52 | 575.32 | 91.54 | .745 | 1.442 |
| Open mindedness by | | | | | |
| Creativity | 605.36 | 592.15 | 98.10 | .881 | 1.220 |
| Tolerance | 621.04 | 605.37 | 111.15 | .797 | .780 |
| Curiosity | 628.48 | 614.18 | 91.13 | .594 | .185 |
| Task performance by | | | | | |
| Persistence | 608.48 | 600.16 | 102.15626 | .610 | .518 |
| Responsibility | 588.50 | 578.72 | 96.22337 | .696 | 1.155 |
| Self-control | 607.07 | 603.97 | 95.05298 | .656 | 1.155 |

**Table 2.**
*Frequencies of groups.*

| Groups | | *N* | *%* |
|---|---|---|---|
| Gender | Girls | 1841 | 58.6% |
| | Boys | 1299 | 41.4% |
| SES | Low | 1046 | 33.3% |
| | Medium | 1047 | 33.3% |
| | High | 1045 | 33.3% |
| SEL Evaluation | No, we don't evaluate these skills | 542 | 17.3% |
| | Yes, using informal evaluation | 1392 | 44.3% |
| | Yes, using formal evaluation | 855 | 27.2% |

**Table 3.**
*Confirmatory factor analysis of big five domain model.*

| Model | $X^2$ | *df* | $x^2/df$ | CFI | TLI | RMSEA | |
|---|---|---|---|---|---|---|---|
| | | | | | | Value | 90% |
| Model 1: 1-main domain, 15 sub-dimension | 4233.838*** | 90 | 47.043 | .650 | .591 | .121 | .118, .124 |
| Model 2: 5-main domain, 15 sub-dimension | 2205.945*** | 80 | 27.574 | .820 | .764 | .092 | .089, .095 |
| Model 3: 5-main domain, 10 subdimension | 575.733*** | 25 | 23.029 | .928 | .871 | .084 | .078, .090 |

*p < .05; **p < .01; ***p < .001

one general factor and in Model 2, 15 subdimensions were hypothesized to be related to five main domains as described in the Big Five Model. Although Model 2 had better fit indices compared to Model 1, CFI, TLI and RMSEA values indicated that the fit was poor. Since the Big Five Model's fit indices showed that it did not adequately match the data, the model was modified by reducing the number of subdimensions in accordance with the lowest factor loadings.

The subdimensions with the lowest factor loadings were eliminated from the big five model. Therefore, tolerance (β = 0.501), self-control (β = 0.695), assertiveness (β = 0.393), trust (β = 0.314), and emotional control (β = 0.619) were removed with the lowest factor loadings in this model. Therefore, the model was modified as five domains and ten subdimensions (Model 3). Fit indices of the modified model indicated acceptable model fit with better CFI, TLI and RMSEA values. Overall, the model provided an acceptable fit (CFI = 0.928, TLI = 0.871, RMSEA = 0.084).

*Predicting Mathematics Achievement*

The modified 5-main domain, 10 subdimensions model of SSES was used to predict the mathematics achievement of students (see Table 4). According to the results, open-mindedness (β = 0.258) and emotional regulation (β = 0.398) main domains showed significant positive relationships with mathematics achievement. On the other hand, engaging with others (β = -0.516) main domain showed significant negative relationship with mathematics achievement. On the other hand, task performance and collaboration main domains did not have a statistically significant relationship with mathematics achievement. Overall, these five main domains explained 8% of the variance in mathematics achievement ($R^2$ = 0.077). Among these domains, engaging with others had the most important role in prediction.

**Table 4.**
*Standardized regression coefficients for predicting mathematics achievement*

| Predictors | Standardized Coefficients | S.E. |
|---|---|---|
| Open mindedness | 0.258*** | 0.053 |
| Task performance | -0.049 | 0.056 |
| Engaging with others | -0.516*** | 0.141 |
| Collaboration | 0.033 | 0.067 |
| Emotional regulation | 0.398** | 0.127 |

*p < .05; **p < .01; ***p < .001

*Differential relationship between SEL and mathematics achievement for related groups*

To evaluate differential relationships, the analyses were conducted again for the gender, SES, and SEL evaluation groups.

**Gender Groups**

To determine whether relationships differ, mathematics achievement for each gender group was predicted using the SSES model. The findings demonstrated that whereas the model explained 6% of the variation in math achievement for boys ($R^2$ = 0.056), it explained 9% of the variation for girls ($R^2$ = 0.092). Thus, the model was able to explain more of the variation of mathematics achievement for girls than for boys.

Even though there was a significant positive relationship between the open-mindedness domain and math achievement for both boys and girls, the association was stronger for boys (β = 0.320) than for girls (β = 0.246). On the other hand, mathematics achievement of girls had a statistically significant positive relationship with emotional regulation (β = 0.412) and a statistically significant negative relationship with engaging with others (β = -0.419), but both domains did not have a statistically significant relationship with mathematics achievement for boys. Furthermore, there is no statistically significant relationship between math achievement for both boys and girls and task performance or collaborative domain.

**Table 5.**
*Standardized regression coefficients in the model for boys and girls.*

| | Mathematics Achievement | Standardized coefficients | S.E. |
|---|---|---|---|
| Boys | Open mindedness | 0.320*** | 0.089 |
| | Task performance | -0.138 | 0.090 |
| | Engaging with others | -0.348 | 0.214 |
| | Collaboration | 0.028 | 0.113 |
| | Emotional regulation | 0.185 | 0.194 |
| Girls | Open mindedness | 0.246*** | 0.063 |
| | Task performance | 0.012 | 0.068 |
| | Engaging with others | -0.419** | 0.152 |
| | Collaboration | -0.064 | 0.075 |
| | Emotional regulation | 0.412** | 0.134 |

*p < .05; **p < .01; ***p < .001

SES Groups

The SSES model was used to predict mathematics achievement for each SES group (low, medium and high) in order to determine whether or not the relationships differ (see Table 6). According to the results, the model explained 13% of the variance in mathematics achievement for students with low SES ($R^2$=0.130), 9% of the variance for students with middle SES ($R^2$=0.091). and 6% of the variation for students with high SES ($R^2$=0.056). Thus, the model explained more variance in mathematical proficiency for students from low SES than for children from medium and high SES.

While there was a statistically significant positive relationship between the open-mindedness domain and mathematical achievement for both low- and high-level SES groups (β = 0.507 and 0.170, respectively), the association was stronger for the former than for the latter. There was a statistically significant negative relationship between mathematical achievement and engaging with others for both low- and high-level SES groups (β = -0.606 and β = -0.471). Finally, only high SES groups showed a statistically significant positive correlation between mathematical achievement and the emotional regulation domain (β = 0.341). There was no significant relationship between any of the domains and math achievement for groups with a medium level SES.

**Table 6.**

*Standardized regression coefficients for low, medium, and high-level SES*

| | Mathematics Achievement | Standardized Coefficients (β) | S.E. |
|---|---|---|---|
| Low SES | Open mindedness | 0.507*** | 0.120 |
| | Task performance | -0.166 | 0.145 |
| | Engaging with others | -0.606* | 0.279 |
| | Collaboration | -0.089 | 0.105 |
| | Emotional regulation | 0.520 | 0.283 |
| Medium SES | Open mindedness | 0.073 | 0.126 |
| | Task performance | -0.108 | 0.239 |
| | Engaging with others | -0.955 | 0.769 |
| | Collaboration | 0.399 | 0.406 |
| | Emotional regulation | 0.691 | 0.640 |
| High SES | Open mindedness | 0.170* | 0.082 |
| | Task performance | 0.104 | 0.080 |
| | Engaging with others | -0.471** | 0.174 |
| | Collaboration | -0.031 | 0.076 |
| | Emotional regulation | 0.341* | 0.149 |

*p < .05; **p < .01; ***p < .001

SEL Evaluation Groups

Mathematics achievement was predicted using the SSES model for the non-evaluated, informally evaluated, and formally evaluated SEL evaluation groups (see Table 7). According to the findings, this model accounted for 6% of the variance in mathematics achievement ($R^2$ = 0.062) for non-evaluated SEL groups, 8% of the variance ($R^2$ = 0.080) for informally evaluated SEL groups, and 9% of the variation for the SEL group that was formally evaluated ($R^2$ = 0.091).

Across all SEL evaluation levels, there was a statistically significant positive relationship between mathematics achievement and open-mindedness domain (β_non = 0.240, β_informally = 0.233, β_formally = 0.359). However, this relationship was getting stronger from non-evaluated groups to formally evaluated groups. Additionally, there were statistically significant negative relationships between the domains of engaging with others and mathematical achievement for groups of students who were formally evaluated (β = -0.482). Additionally, for this group, the emotional control domain and mathematical achievement were positively correlated (β = 0.343). Achievement in mathematics did not exhibit a statistically significant relationship with other domains in the non-evaluated and informally evaluated groups.

## Discussion

The primary objective of this study is to examine the extent to which the Big Five domains of social and emotional skills can predict mathematics achievement among high school students in Turkey. Specifically, the study aims to explore whether the predictive power of these domains is moderated by factors such as gender, SES, and the level of SEL evaluation. This investigation is premised on the understanding that social and emotional competencies are crucial not only for personal well-being and social integration but also for academic success. By focusing on these competencies, the study seeks to provide a more nuanced understanding of how SEL is related to mathematics achievement, offering potential insights into how educational strategies can be tailored to meet the diverse needs of students (CASEL, 2020; OECD, 2021). The current study is significant not only because it applies the Big Five model to a Turkish context but also because it provides a detailed analysis of how SEL might influence mathematics achievement across different student groups. The findings have the potential to contribute to the broader literature on SEL by offering new insights into how social and emotional competencies interact with gender, SES, and school-level evaluation practices to shape academic outcomes. Moreover, the study's focus on mathematics achievement is particularly relevant given the global emphasis on STEM (Science, Technology, Engineering, and Mathematics) education as critical for future workforce development (OECD, 2021).

According to the results of the current study, the social and emotional skills domains accounted for 8% of the variance in mathematics achievement, particularly with open-mindedness, emotional control, and engaging with other factors. Thus, social and emotional skills have been shown to support academic achievement, which is consistent with previous research findings (Chernyshenko et al., 2018; McCormick et al., 2015; OECD, 2021).

The domains of emotional regulation and open-mindedness are positively related to students' proficiency in mathematics. Accordingly, the model suggests that students who have gained emotional control and an open mind typically perform better in math classes. The results align with the OECD's SSES report for every city that took place (OECD, 2021). Open-mindedness domain which was defined in the study as openness to experience with the combination of skills, curiosity, and creativity has significantly positive relationship with students' mathematics performance (Eroğlu et al, 2021; OECD, 2021). As a result, students who were classified as extremely creative and curious also said they were willing to learn new things, which leads to better academic performance (OECD, 2021). Additionally, there is a favorable correlation between mathematical achievement and the emotional regulation domain, which was defined as emotional stability with the combination of skills, stress resilience, and optimism (CASEL, 2020; Eroğlu et al, 2021). As a result, students who have mastered emotional regulation are more likely to perform better in mathematics.

On the other hand, engaging with others plays the most important role in predicting mathematics achievement in the sample data from Turkey. The outcome is in line with the OECD's SSES findings for the older group across all data. More social 15-year-old children receive lower math grades, according to the data. Individuals experience physiological and physical changes and are impacted by their peers during adolescence (Ahmetoglu, 2009; Gander & Gardiner, 2004). Teenagers' top priority during this time is getting their peers to accept them (Durualp, 2014). The study's conclusions might thus be linked to the reality that teens prioritize their relationships and social ties over academic success (OECD, 2021).

*The differential relationship between SEL and mathematics achievement*

One of the objectives of the current study was to use the big five domains model of SSES to predict mathematics achievement for various groups to investigate differential relationships. Having a general finding may not apply to every subgroup, thus the differential relationships offer a deeper understanding of a phenomenon, which is necessary to get better insight and use the findings efficiently. In order to examine differential relationships, the current study used gender, SES, and SEL evaluation level as subgroups.

The findings showed that explained variance of girls in math achievement was greater than boys. As a result, social and emotional competencies and mathematical achievement of girls were more correlated than boys. Previous research stated that gender differences have a key role in girls' better development of social and emotional skills compared to males' (Durualp, 2014; Kabakci & Korkut, 2010; Memis & Memis, 2013). Compared to boys, girls are found to have superior communication abilities and behaviors, including starting a discussion, adjusting, sustaining interaction, and being emotionally sensitive (Durualp, 2014; Kabakci & Korkut, 2010). Furthermore, mathematics performances of boys and girls who have higher curiosity and creativity skills are more likely to become more developed. On the contrary, more sociable and energetic girls are likely to have lower mathematics scores while boys' sociability and energy skills are not related to their mathematics performance.

The results showed that the explained variance in mathematical achievement was 13% for students from low socioeconomic backgrounds, 9% for students from medium socioeconomic backgrounds, and 6% for students from high socioeconomic backgrounds. Thus, it can be said that the social and emotional skills predicted math achievement more accurately for students from lower socioeconomic backgrounds. Therefore, students from disadvantaged socioeconomic backgrounds are likely to perform better academically in mathematics if they have acquired social and emotional skills. On the other hand, their performance in mathematics tends to suffer if they lack social and emotional skills. Thus, the current study shows how important it is to help pupils from low socioeconomic backgrounds.

According to the findings, this model accounted for 6% of the variation in math achievement for SEL groups that were not evaluated, 8% of the variation in math achievement for SEL groups that were evaluated informally, and 9% of the variation in math achievement. Thus, it can be said that the SSES model and mathematics achievement for evaluated SEL groups and non-evaluated groups have differential relationships. The findings supported the literature's assertion that SEL assessment and evaluation are critical components of the development of these abilities, which are linked to academic success (Agliati et al., 2020; CASEL, 2019; Sutton, 2021).

By providing empirical evidence on the relationship between SEL and mathematics achievement, the current study can inform the design and implementation of more effective SEL programs in Turkish schools. Additionally, it offers valuable

insights for policymakers and educators looking to enhance educational outcomes through targeted interventions that address the diverse needs of students. Ultimately, the study's findings could contribute to the development of a more equitable and effective educational system that supports the holistic development of all students, regardless of their gender, socioeconomic background, or the extent to which SEL is formally evaluated in their school.

**Limitations**

The study has limitations. One of the important limitation is that the current study was carried out using secondary data that was gathered by the OECD. Thus, the results from the study, which included students in the ninth, tenth, and eleventh grades in the Turkish sample, did not support the major five domains model of the OECD's SSES. As a result, the model was adjusted to exclude the subdimensions with the lowest factor loadings. For this reason, the model was examined using fewer subdimensions.

**References**

Agliati, A., Benitez, I., Cavioni, V., & Elisabetta, C. (2020). *Toolkit for assessing social and emotional skills at school*. Lithuanian Children and Youth Centre.

Furnham, A., Monsen, J., & Ahmetoglu, G. (2009). Typical intellectual engagement: Big Five personality traits, approaches to learning and cognitive ability predictors of academic performance. *British Journal of Educational Psychology, 79*(4), 769–782. https://doi.org/10.1348/978185409X412147

Assessment Work Group. (2019). *Student social and emotional competence assessment: The current state of the field and a vision for its future*. Collaborative for Academic, Social, and Emotional Learning.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Lawrence Erlbaum Associates, Inc.

CASEL. (2019). What is SEL? https://casel.org/what-is-sel/CASEL. (2020). CASEL's SEL framework: What are the core competence areas and where are they promoted? https://casel.org/casel-sel-framework-11-2020/

Chernyshenko, O., Kankaraš, M., & Drasgow, F. (2018). Social and emotional skills for student success and well-being: Conceptual framework for the OECD study on social and emotional skills. *OECD Education Working Papers, No. 173*. OECD Publishing. https://doi.org/10.1787/db1d38e3-en

Durlak, J. D., Domitrovich, C. E., Weissberg, R. P., & Gullotta, T. P. (2015). *Handbook of social and emotional learning: Research and practice*. The Guilford Press.

Durlak, J. A., Weissberg, R. P., Dymnicki, A., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*(1), 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

Durualp, E. (2014). Ergenlerin sosyal duygusal öğrenme becerilerinin cinsiyet ve sınıfa göre incelenmesi. *The Journal of Academic Social Science Studies, (26)*, 13–25.

Eroğlu, E., Suna, E., Taşkireç, B., & Yaşaran, Ö. Ö. (2021). OECD sosyal ve duygusal beceriler araştırması Türkiye ön raporu. *Eğitim Analiz ve Değerlendirme Raporları Serisi, No. 19*. T.C. Millî Eğitim Bakanlığı.

Gander, M. J., & Gardiner, H. W. (2004). *Çocuk ve ergen gelişimi* (B. Onur, Çeviri ed.). İmge Kitabevi. (E.T. 10.08.2016)

Gol-Guven, M. (Ed.). (2021). *Çocuklukta sosyal ve duygusal öğrenme*. Yeni İnsan Yayınevi. Istanbul

Green, A. (2024) *Artificial intelligence and the changing demand for skills in the labour market*. OECD Artificial Intelligence Papers, No. 14, OECD Publishing, Paris, https://doi.org/10.1787/88684e36-en

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114–158). The Guilford Press.

Kabakci, Ö. F., & Korkut, F. (2010). 6- 8. sınıftaki öğrencilerin sosyal-duygusal öğrenme becerilerinin bazı değişkenlere göre incelenmesi. *Eğitim ve Bilim, 33*(148), 77–86.

Kankaraš, M., & Suarez-Alvarez, J. (2019). Assessment framework of the OECD study on social and emotional skills. *OECD Education Working Papers, No. 207*. OECD Publishing. https://doi.org/10.1787/734083b3-en

McCormick, M. P., Cappella, E., O'Connor, E. E., & McClowry, S. G. (2015). Social-emotional learning and academic achievement: Using causal methods to explore classroom-level mechanisms. *AERA Open, 1*(3). https://doi.org/10.1177/2332858415603959

Memis, A., & Memis, U. A. (2013). Gender, achievement and social skill. *Karaelmas Journal of Educational Sciences, 1*(1), 43–49.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Muthén & Muthén.

OECD (2021). *Beyond academic learning: First results from the survey of social and emotional skills*. OECD Publishing. https://doi.org/10.1787/2ed1b046-enOECD. (2021). *OECD survey on social and emotional skills technical report*. OECD Publishing. https://doi.org/10.1787/3f50b556-en

OECD (2024). *Nurturing Social and Emotional Learning Across the Globe: Findings from the OECD Survey on Social and Emotional Skills 2023*. OECD Publishing, Paris, https://doi.org/10.1787/32b647d0-en.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322–338. https://doi.org/10.1037/a0014996

Rice, C., & Pasupathi, M. (2010). Reflecting on self-relevant experiences: Adult age differences. *Developmental Psychology, 46*(2), 479–490. https://doi.org/10.1037/a0018098

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. Lawrence Erlbaum Associates.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Sutton, E. (2021). How to measure SEL - 7 approaches to consider. https://www.branchingminds.com/

Taylor, R., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child Development, 88*(4), 1156–1171. https://doi.org/10.1111/cdev.12864

Ullman, J. B. (2001). Structural equation modeling. In B. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (4th ed., pp.653-771). Allyn & Bacon

This page is intentionally left blank.

www.iejee.com

iejee

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

# Improving Context Scale Interpretation Using Latent Class Analysis for Cut Scores

Liqun Yin[a], Ummugul Bezirhan[b], Matthias von Davier[c]

Received : 6 January 2025
Revised : 23 January 2025
Accepted : 2 March 2025
DOI : 10.26822/iejee.2025.378

[a] **Corresponding Author:** Liqun Yin, TIMSS & PIRLS International Study Center, Boston College USA.
E-mail: yinld@bc.edu
ORCID: https://orcid.org/0009-0005-1919-3548

[b] Ummugul Bezirhan, TIMSS & PIRLS International Study Center, Boston College, USA.
E-mail: bezirhan@bc.edu
ORCID: https://orcid.org/0000-0002-8771-4780

[c] Matthias von Davier, TIMSS & PIRLS International Study Center, Boston College, USA.
E-mail: vondavim@bc.edu
ORCID: https://orcid.org/0000-0003-1298-9701

## Abstract

This paper introduces an approach that uses latent class analysis to identify cut scores (LCA-CS) and categorize respondents based on context scales derived from large-scale assessments like PIRLS, TIMSS, and NAEP. Context scales use Likert scale items to measure latent constructs of interest and classify respondents into meaningful ordered categories based on their response data. Unlike conventional methods reliant on human judgments to define cut points based on item content, model-based approaches such as LCA find statistically optimal groups, a categorical latent variable, that explains item score differences based on score distribution differences between latent classes. Cut scores for these classes are determined by conditional probability calculations that relate class membership to observed scores, finding the intersection point of adjacent smoothed probability distributions and connecting it to the construct. Demonstrated through application to PIRLS 2021 data, this is useful to validate existing categorizations of the context scale by human experts, and can also help to enhance classification accuracy, particularly for scales exhibiting highly skewed distributions across diverse countries. Recommendations for researchers to adopt this LCA-CS approach are provided, demonstrating its efficiency and objectivity compared to judgment-based methods.

**Keywords:**

Context Scales, Latent Class Analysis, Cut Scores, Large-scale Assessments

## Introduction

In educational assessments of achievement, standard-setting has been used for meaningful interpretation of test scores and for making decisions that impact students' educational trajectories, such as screening students for instruction, grade promotion, selection, or admission (e.g., Cizek & Bunch, 2007; Cizek, 2012; Jiao et al., 2011). Performance standards, which are set through carefully determined cut scores, serve to classify examinees into defined proficiency levels, in doing so, guiding stakeholders' understanding of individuals' competencies relative to a given domain (Cizek, 2012). Therefore, standard-setting is central to establishing that assessments function not only as measurement tools but also as benchmarks for educational quality and progress.

Traditionally, standard-setting methods implemented for achievement instruments have relied on subject matter experts (SMEs) to interpret the content of assessment items and determine cut scores that align with descriptions of performance levels (Cizek, 1993). These methods are generally categorized as test-centered, where SMEs focus on individual test items, or examinee-centered, where judgments are based on examinee performance rather than specific item content (Jaeger, 1989). Methods such as the Angoff procedure (Angoff, 1971), bookmark method (Mitzel et al., 2013), and contrasting groups method (Livingston & Zieky, 1989) are widely used in standard-setting. In these approaches, SMEs discuss the difficulty of test items and the expected performance of a "borderline" examinee to set a threshold for each proficiency level (Cizek, 2005; Peabody et al., 2023). The Angoff and bookmark methods are test-centered, as they focus on the properties of individual test items, with SMEs evaluating item difficulty to estimate the performance of a minimally competent examinee. In contrast, the contrasting groups method is examinee-centered, as it relies on SMEs classifying examinees directly based on their overall performance relative to the standard.

In addition to test-centered and examinee-centered distinctions, standard-setting methods can be classified as holistic or analytical, norm-referenced, or criterion-referenced. Holistic methods involve evaluating overall performance levels, while analytical methods break down performance into specific competencies or skills. Norm-referenced methods set performance standards by comparing the examinee's performance to a reference group, whereas criterion-referenced methods define standards based on specific performance criteria or competencies (Cizek, 2012). Similar to the test-centered versus examinee-centered distinction, these categorizations, while conceptually useful, tend to overlap in practice, as most standard-setting approaches combine elements of various methodologies to comprehensively evaluate examinee proficiency levels.

Although well-established, these methods require intensive cognitive effort from experts to consider both the test content's characteristics and the abilities of the target population. They are susceptible to inconsistencies due to variations in judgment, especially across diverse contexts (Brown, 2007; Cizek, 2012).

To address the limitations of traditional judgment-based approaches, recent research has explored data-driven methods for setting cut scores (Binici & Cuhadar, 2022; Brown, 2007; Peabody et al., 2023; Templin & Jiao, 2012). Latent class analysis (LCA; Dayton & MacReady, 1976, 2006; Lazarsfeld & Henry, 1968) has emerged as an appealing alternative for establishing cut scores in a statistically objective manner. LCA, a categorical latent variable modeling technique, identifies groups within a population based on response patterns rather than judgment, thus reducing the subjectivity typically associated with standard setting. This approach segments examinees into homogeneous latent classes according to a statistical optimization criterion, effectively distinguishing groups based on the item response distributions within each class. Unlike conventional methods that presuppose a continuous latent trait, LCA models assume that different, discrete latent classes account for variation in observed scores. This enables LCA to categorize individuals into performance levels based on empirical relationships among responses rather than a-priori content-based judgments.

Brown (2007) evaluated the effectiveness of LCA alongside the Angoff procedure and profile rating method for a middle school statistics assessment. This study utilized LCA to categorize students based on response patterns, providing an empirical, data-driven alternative to judgment-based approaches. The results showed that the traditional methods showed strong agreement, with students categorized similarly 85.7% of the time. The LCA showed an even higher alignment with the Angoff method (92.2%) but slightly lower agreement with the Profile method (77.1%), indicating that LCA could reliably classify students into proficiency levels while reducing reliance on expert judgment. Similarly, Binici and Cuhadar (2022) applied LCA to an operational large-scale science assessment administered in one of the southern states in the United States to validate performance standards derived from traditional methods. Their work examined whether LCA could provide additional validity evidence into the classification accuracy of existing cut scores. By analyzing the latent structure within student response patterns, Binici and Cuhadar (2022) demonstrated that LCA could complement conventional judgment-based methods by offering a statistically derived basis for performance standards. These studies showcase the advantages of LCA in creating objective and data-driven cut scores, primarily focusing on setting performance standards for achievement data. While applying LCA to standard settings is not entirely new, its application to background scales remains relatively underexplored.

In large-scale international assessments such as PIRLS (Progress in International Reading Literacy Study) and TIMSS (Trends in International Mathematics and Science Study), context questionnaires are widely used to gather data on students' background through student, school, and home questionnaires. Many of these context items are designed to measure common and dominant underlying latent constructs, such as student motivation, family support, and school resources, which aid in understanding the various factors that relate to student performance. The item

response theory (IRT) based scaling approach is then utilized to derive context scale scores for the items measuring the same latent construct.

In operational settings, context scales are often divided into regions aligned with raw score points and transformed reporting scale cut points. The interpretation of these regions is content-referenced, meaning that each boundary aligns with a combination of response categories. These cut points are often defined through SME judgments. Hence, experts determine what constitutes high or low levels on each scale, sometimes solely based on reviewing the items and response categories, without referencing how respondents use the scale. However, these content-referenced cut-score definitions can result in score regions that contain few or no students, especially when evaluating skewed scale distributions across countries with diverse educational backgrounds.

Current study introduces an LCA-based cut score (LCA-CS) determination approach that addresses the limitations of traditional, judgment-based cut score definitions on context scales. This approach uses LCA with a predefined number of classes determined as the number of ordered categories experts wish to distinguish. LCA identifies groups of examinees based on their observed responses, providing posterior probabilities of class membership for each individual. Examines are then assigned to the most likely class based on the maximum posterior class probability, therefore classifications are statistically grounded rather than subjective expert judgment. After LCA identifies latent classes, which are homogeneous groups within the data, the latent classes are sorted based on the expected mean score for each class. This step reflects the principles of located and ordered latent class models (Clogg 1979; Croon, 1990; Formann 1992; Lazersfeld & Henry, 1968) that the classes are represented by scores on a latent continuum. In our case, the construct's scale score provides this continuum, ensuring that class order is directly related to the underlying latent trait. This can be interpreted as the probability of selecting increasingly positive categories on a rating scale, in the case of context scales, or for cognitive skills, selecting the correct response, which increases as one progresses through a set of latent classes from the lowest to the highest (Croon, 2002), making it particularly useful in contexts where subgroups within a latent trait are to be identified rather than measuring differences between individuals. However, for ordered latent class approach to hold, it is also necessary to verify that the expected scores follow the same order across all items. Additionally, the differences between the expected scores for adjacent classes should be sufficiently large to demonstrate meaningful separation.

Furthermore, we modeled the conditional score distributions for each class independently to identify cut scores that separate adjacent classes. For this, we assume that each latent class represents a homogeneous group, and the conditional distribution of scores within each class follows a normal distribution. The use of conditional normal approximations for score distributions reflects widely applied practices in latent variable modeling, where parametric assumptions are employed to smooth score distributions (e.g., Heinen, 1993, 1996; Embretson & Riese, 2013; Mislevy, 1983; Rost & von Davier, 1995; Smit et al., 2003; Templin & Jiao, 2012). While Formann (1992) emphasizes the relationship between categorical latent variables and response probabilities in linear logistic latent class models, our model ties class membership to a latent continuum. Smoothing these distributions helps cut-score boundaries not to be overly sensitive to random fluctuations in the data. This is particularly important in large-scale assessments where sample sizes and response patterns vary widely across contexts.

When applying LCA, the intersection points of smoothed posterior probabilities between adjacent classes define the cut scores. Then, these cut points are mapped back to the underlying construct. The model integrates categorical class definitions with continuous construct measurement by anchoring these cut scores to the IRT scale. Templin and Jiao (2012) argue for combining latent class models with continuous scaling to enhance the psychometric validity of classifications, while Rost (1990) emphasizes the compatibility of latent class and trait models for defining ordered categories along a latent continuum. Similarly, Croon (1990) and Formann (1992) offer theoretical frameworks for modeling ordered latent classes that align with continuous latent constructs, providing a basis for statistically grounded and construct-aligned classifications. Leveraging these principles, our approach bridges the strengths of LCA and IRT to develop a replicable, robust, and easy-to-implement method for cut-score determination, making the classification more apt for secondary analysis and interpretation of the results.

To demonstrate the applicability of our model, we utilize PIRLS 2021 data to validate the classifications on context scales and enhance classification accuracy, particularly for scales with skewed distributions across countries with diverse educational backgrounds. This data-driven approach strengthens examinee categorization by extending the application of LCA-based approaches to standard setting and proficiency scaling into new domains, supporting reliable, data-driven standard setting across different educational contexts. Overall, this study highlights the advantages of LCA-CS as a viable alternative or complementary method to traditional judgment-based approaches for determining cut scores on context scales.

## Methods

The latent class model (e.g., Lazarsfeld & Henry, 1968; von Davier & Lee, 2019) is a statistical technique for identifying latent subgroups within a population based on categorical observed variables. Suppose we observe J polytomous items *(j=1,2,...,J)* where each item has $K_j$ *(k = 1,...,$K_j$)* response categories, and we observe responses for examinees *i* = 1,2,...,N. The observed responses to these variables are denoted as $X_{ijk}$, where $X_{ijk}$ = 1 if examinee i selects the k-th response category to the j-th item, and 0 otherwise. The latent class model assumes that the observed joint distribution of the manifest variables can be expressed as a weighted sum of conditional distributions in C latent classes. Each class represents a cross-classification table of response probabilities, parameterized by $\pi_{jck}$, the probability of selecting the *k*-th response to the *j*-th item in class *c*. For each variable *j*, $\sum_{k=1}^{K_j} \pi_{jck}$ = 1. The weights $p_c$, referred to as the mixing proportions, represent the prior probabilities of class membership satisfying $\sum_{c=1}^{C} p_c$ =1.

A key assumption in LCA is conditional independence, meaning that the observed variables are independent of one another, given membership in a latent class. This assumption, analogous to the local independence property in IRT, allows the model to decompose the observed joint distribution of responses into class-conditional probabilities (Yamamato, 1987). The model is fully identified by the matrix of conditional probabilities, $\pi_{jck}$, and the class distribution, $p_c$ which together parameterize the probability of observed responses.

Under conditional independence, the probability of observing a specific set of responses for an individual *i* in a class *c* is given by:

$$f(X_i;\pi_c) = \prod_{j=1}^{J}\prod_{k=1}^{K_j}(\pi_{jck})^{X_{ijk}}$$

The probability of the observed responses across all classes is then

$$P(X_i|\pi,p) = \sum_{c=1}^{C} p_c \prod_{j=1}^{J}\prod_{k=1}^{K_j}(\pi_{jck})^{X_{ijk}}$$

The parameters of the model $p_c$ and $\pi_{jck}$ are estimated by maximizing the log-likelihood function:

$$lnL = \sum_{i=1}^{N} ln\left(\sum_{c=1}^{C} p_c \prod_{j=1}^{J}\prod_{k=1}^{K_j}(\pi_{jck})^{X_{ijk}}\right)$$

Posterior probabilities for class membership are computed using Bayes' rule:

$$P(c|X_i) = \frac{p_c f(X_i;\pi_c)}{\sum_{q=1}^{C} p_q f(X_i;\pi_c)}$$

where *c* = 1,2,...,C.

Latent classes are ordered if there is a permutation $\eta(c)$ of the class membership variable C so that the expected responses of all items *j* are ordered across classes. That is,

$$\sum_{k=1}^{K_j} k\pi_{j[\eta(c)]k} < \sum_{k=1}^{K_j} k\pi_{j[\eta(c)+1]k} \quad \text{for all } j = 1,..,J$$

This ensures that an ordered or continuous latent trait that leads to equivalent conditional probabilities can be identified. To test this, the classes are ordered by their expected sum score, i.e., the expected score is increasing with (reordered) class index. Then, the same property, the monotonicity of the expected scores, is checked for each item on the scale (Rost, 1990).

### LCA for Identifying Cut Points

The proposed approach uses the latent class model to identify cut points on a scale from the response data. It first uses LCA to define a categorical latent variable that explains differences in item scores based on score distribution differences between homogeneous groups (latent classes). Next, a series of calculations are needed to identify cut points on the context scale. The details of these steps are described below. The following descriptions are based on three classes for simplicity and clarity, though the procedure generalizes to any number of classes.

1. Run latent class analysis (LCA) with a pre-specified number of classes. This number is usually identified based on literature or by context experts. In large-scale assessments such as TIMSS & PIRLS, the goal is to define cutpoints for three groups with high, medium, and low expected scores on the context scales.

2. Assign test takers to classes based on the posterior probability $P(C = c|X_1...X_J)$ of being a member of class c given responses $X_1...X_j$ to a set of items. Each test taker is assigned to the class based on the maximum posterior probability among the specified classes.

3. Re-order classes so that the expected score increases with the class index. That is, *E(score|C=c) > E(score|C=c+1)* if class *c = 1* represents the class with higher scores, where *E(score|C=c)* is the expected score given class c. Meanwhile, check whether the expected scores of each item are in the same order as the ordered classes.

4. Calculate the probability of a score given a class, *P(score|C)*. This probability is approximated assuming that each class is a

homogeneous group with a conditional normal ability distribution, $N(\mu_c,\sigma_c)$, where $\mu_c$ and $\sigma_c$ are the mean and standard deviation of scores within the class. The result is an approximate conditional probability distribution, the probability of a score given a class, *P(score|C)*.

5. Calculate the conditional probability approximation of a "class" given a score using Bayes' theorem. Standard results yield,

$$P(C|score) = \frac{P(score|C)P(C)}{P(score)}$$

where *P(score|C)* is obtained from step 4. *P(C)* is the class size, and *P(score)* is the marginal probability for each score point.

6. Identify the cut score points and connect them to the construct, either the raw points or the scale score. The cut points are identified by locating the intersection point of adjacent smoothed posterior probability distributions, obtained from step 5, so that *P(C=c|cut point) > P(C=c+1|cut point)* and *P(C=c+1|cut point-1>) P(C=c|cut point-1)*, if class *c* = 1 represents the class with higher scores.

7. Classify the respondents into one of the three regions based on the identified cut points. Once the cut points are determined using this method, the subsequent procedures of assigning respondents to categories mirror those of the judgment-based cut point specification method or other methods.

For reporting or interpretation of the regions divided by these cut points, the minimum responses needed to meet or exceed the cut scores could be determined by calculating the expected responses for each item based on the IRT model and estimated item parameters. This involves selecting the most likely response for each item given the associated scale cut score, starting with the response category with the highest probability across all items, then moving to the next highest probability on another item until the total raw scores of expected responses are achieved to have the same values as the identified raw cut scores. Note that any response pattern that matches the raw score associated with the scale cut score is compatible with this approach if the scale score is derived using Rasch IRT model, just as in the judgement-based approach.

## Application of the LCA-CS Method for Creating Scale Regions

### *PIRLS and Context Scales Reporting*

This section describes applying the approach to define scale regions using data from PIRLS 2021. PIRLS is designed to measure reading achievement at the fourth-grade level and school and teacher practices related to reading instruction. Students complete a reading assessment and a questionnaire asking about their attitudes toward reading and reading habits. In addition, parents, teachers, and school principals are given questionnaires to gather information about students' home and school experiences in developing reading literacy. Since 2001, PIRLS has provided high-quality data for monitoring progress in students' reading achievement in their fourth year of schooling and measuring trends in achievement over time, covering 20 years of trends.

In PIRLS 2021, the fifth assessment cycle, 57 countries and 8 benchmarking entities participated. All students were administered the same questionnaires after the achievement booklet administration. PIRLS 2021 collected data from approximately 400,000 students, their parents, teachers, and school principals (Mullis et al. 2023). The PIRLS context questionnaire included several item sets intended to measure a latent construct. These constructs included the availability of home resources for learning, participation in literacy and numeracy activities in the home, the school's emphasis on academic success, students' attitudes about learning, and many others. In total, 22 context scales were derived from the PIRLS 2021 data collected from students, their parents, teachers, or principals using the Rasch partial credit model (PCM; Masters, 1982; Masters & Wright, 1997). The estimated Rasch scale scores were converted into a (10, 2) reporting metric for each scale, based on the countries included in the calibration (Yin & Reynolds, 2023). The reporting metric of the scale is set during the PIRLS cycle when the scale is first used or if a scale was revised by adding or changing items or revising response options.

Respondents were classified into three regions corresponding to high, middle, and low values on the construct to facilitate interpretation of the context scale results. The cut scores on the scale delimiting the regions were described in terms of combinations of response categories, the score combinations needed to reach medium or high score regions were defined based on review by content experts. Details on this procedure can be found in Yin & Reynolds (2023).

Once the raw cut points were identified, the corresponding scale cut scores were located utilizing the fact that the raw score is a sufficient statistic in the Rasch model (Andersen, 1977). This conversion was done assuming all questions in the set were answered. This judgment-based method works well under certain conditions, and the scale is well-centered and has sufficient variance along the range of possible scores. However, when the item responses are highly skewed across countries, the content-referenced cut-score definitions might produce score regions that do not contain students for some reporting groups, or even in some countries. The classification is not very useful, if, for example, only 'medium' and 'high' groups are populated, but no students are assigned to the 'low' group. For analytic purposes, such a case would reduce the reporting to only two groups.

In these cases, the proposed LCA-based approach can improve the situation. In the example of the PIRLS 2021 data, the proposed LCA-CS method validates existing categorizations on the context scales and enhances classification accuracy, particularly for scales exhibiting highly skewed distributions across diverse countries.

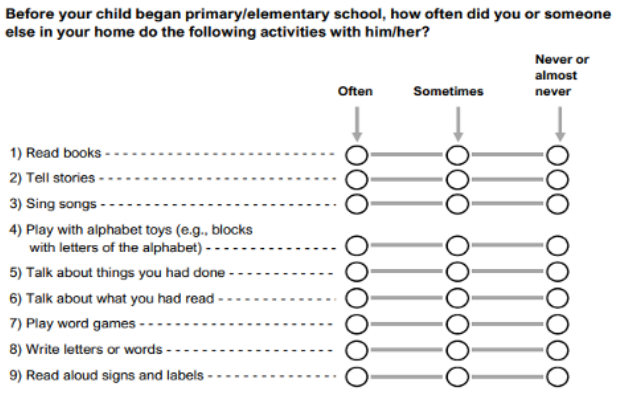
*Description of Example Scale*

This study uses the "Home Early Literacy Activities Before Primary School" scale as an example to demonstrate the LCA-CS method for specifying cut points.

The Home Early Literacy Activities scale was initially developed in PIRLS 2011 and has been continued for subsequent cycles. It includes nine component items from parents' questionnaires, focusing on how often parents engage their children in early literacy activities, as listed in Table 1 (Mullis et al., 2023). All 9 questions have three response options, "Often', "Sometimes", and "Never or almost never", with assigned numeric values of 2, 1, and 0 to the corresponding response categories. Therefore, the maximum available total raw points of this scale were 18.

**Table 1:**

*Questions Included in PIRLS 2021 Home Early Literacy Activities Before Primary School Scale*




The distribution of this scale is highly skewed, with almost no respondents falling into the low category for most countries when using cut scores provided by content experts. The categorization was based on scale cut scores of 10.7 and 6.2, derived from raw cut points of 14 and 4 based on minimal response profiles provided by content experts described earlier.

*Applying the LCA-CS Method*

To apply the proposed LCA-CS method for identifying the raw cut points, the SAS procedure PROC LCA (Lanza et al., 2015), one specialized function designed for latent class analysis in SAS program, was used for estimating the latent class model. The LCA was based on the combined data from all 40 calibration countries (Yin & Reynolds, 2023), countries that administered the assessment as scheduled at the end of the 4th school year, with complete responses to the 9 items. A total of 171,796 respondents were included in the LCA model, estimated assuming three classes to align with the reporting goals for PIRLS 2021 international results. The NSTARTS value in PROC LCA was set as 20 to find the best estimates and avoid local maxima of the likelihood function when conducting the analysis.

The posterior probability of the three classes for each respondent is part of the derived statistics that can be obtained through the SAS LCA procedure. Next, the rest of the steps from the previous section were applied. Table 2 shows the results after step 5, the re-calculated conditional probability approximations of the three classes given a score, *P(C|score)*. In the table, class 1 represents the class with the highest expected score, while class 3 represents the class with the lowest expected score. The left two columns are raw possible total points of complete responses of nine items and the associated unique transformed Rasch scale scores, which were retrieved from Appendix 15B in the PIRLS 2021 context scaling chapter (Yin & Reynolds, 2023). The last three columns are the conditional probability approximations, or smoothed posterior probabilities, for the three classes.

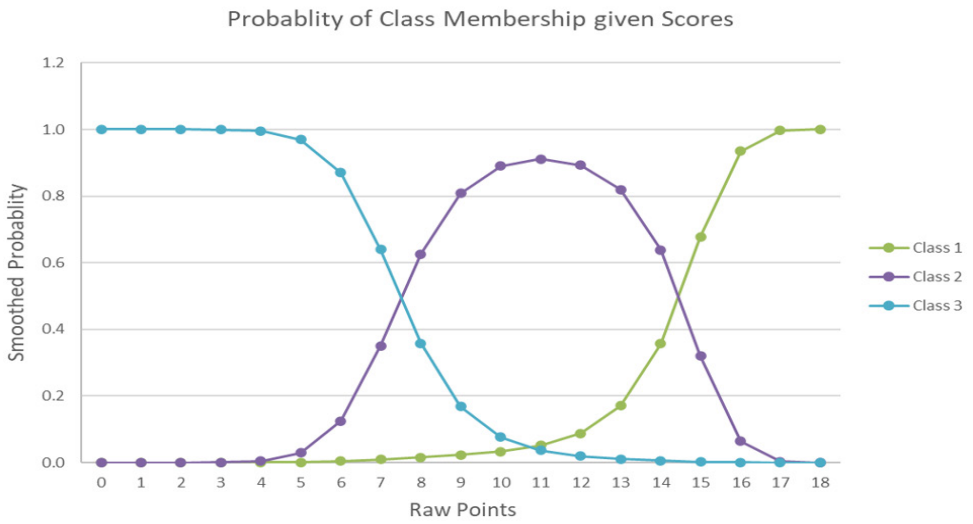
**Table 2:**

*Conditional Probability Approximations of Classes given a Raw Score Point*

| Raw Points | Scale Score | Number of respondents | Smoothed Conditional Probability | | |
|---|---|---|---|---|---|
| | | | Class 1 | Class 2 | Class 3 |
| 0 | 2.0717 | 511 | 0.00 | 0.00 | 1.00 |
| 1 | 3.9169 | 402 | 0.00 | 0.00 | 1.00 |
| 2 | 4.8778 | 698 | 0.00 | 0.00 | 1.00 |
| 3 | 5.5848 | 987 | 0.00 | 0.00 | 1.00 |
| 4 | 6.1700 | 1566 | 0.00 | 0.00 | 1.00 |
| 5 | 6.6863 | 2470 | 0.00 | 0.03 | 0.97 |
| 6 | 7.1652 | 3537 | 0.00 | 0.13 | 0.87 |
| 7 | 7.6184 | 5215 | 0.01 | 0.35 | 0.64 |
| 8 | 8.0567 | 7310 | 0.02 | 0.63 | 0.36 |
| 9 | 8.4885 | 12000 | 0.02 | 0.81 | 0.17 |
| 10 | 8.9179 | 12752 | 0.03 | 0.89 | 0.08 |
| 11 | 9.3525 | 15367 | 0.05 | 0.91 | 0.04 |
| 12 | 9.7989 | 18101 | 0.09 | 0.89 | 0.02 |
| 13 | 10.2674 | 19713 | 0.17 | 0.82 | 0.01 |
| 14 | 10.7707 | 19484 | 0.36 | 0.64 | 0.01 |
| 15 | 11.3376 | 17865 | 0.68 | 0.32 | 0.00 |
| 16 | 12.0220 | 14082 | 0.94 | 0.06 | 0.00 |
| 17 | 12.9578 | 9721 | 1.00 | 0.00 | 0.00 |
| 18 | 14.7746 | 10015 | 1.00 | 0.00 | 0.00 |

Figure 1 displays the smoothed posterior probability distribution for each class. Cut points were identified by locating the intersections of adjacent probability distributions and connecting them to the construct. From Figure 1, the intersections occur between 7 and 8 for classes 2 and 3, and between 14 and 15 for classes 1 and 2. To align with the judgment-based raw cut points approach using whole numbers, 8 and 15 were chosen as the raw cut points.

**Figure 1:**
*Plot of the Conditional Probability Approximations of Classes given a Raw Score Point*




Once the raw cut points were determined, the subsequent procedures of assigning respondents to categories mirror those of the judgment-based cut point specification method described in creating the PIRLS 2021 context scales chapter (Yin & Reynolds, 2023). According to the equivalence table of the raw scores and transformed scale scores presented in Table 2, the corresponding scale scores are 8.0567 and 11.3376 for raw points 8 and 15, respectively. Following the same rounding rules as the judgment-based cut point specification methods, the rounded scale scores, 8.1 (rounded up) and 11.3 (rounded down), were the final scale cut scores. These two cut scores were then used to classify all the respondents into one of three regions, including those from the countries with delayed administrations due to pandemic-related delays.

*Categorization Results Using the LCA-CS Method*

The following section presents the categorization results applied to the Home Early Literacy Activities scale using the LCA-CS method to identify the cut scores.

Table 3 shows the percentage of students whose parents were classified into each of the three regions using two different categorization methods. The standard errors (SEs) associated with the percentages, except for the percentage of 2 or smaller, are listed in parentheses. This table reports the results based on all PIRLS 2021 countries with comparable data, including those not included in the LCA model and item calibrations. The rightmost column shows each country's average scale score and associated SE. The results in the left part of Table 3 are the PIRLS 2021 published results (Mullis et al., 2023), showing percentages derived from conventional methods reliant on human judgments to define raw cut points based on item content. In contrast, the percentages for the three regions in the right part of the table were obtained using the LCA-CS procedures.

In Table 3, within the low region of the scale, there are many very small percentages, 2%, 1%, and even 0s, when using the judgment-based categorization. In practice, reporting the achievement levels for such a small percentage of students in a region is associated with a large error, and PIRLS does not report groups smaller than 2% in size. Therefore, the results from this categorization provided limited value for interpreting the relationship between achievement and home early literacy activities.

In contrast, using the LCA-CS procedures, the distribution of percentages across the three regions is less skewed across countries, enhancing the interpretation of the achievement and the related context. Based on the categorical latent variable modeling technique, the low category is no longer empty for all countries, which identifies groups based on a statistically optimal criterion. Additionally, the percentages in the middle region closely align with those from the judgment-based approach at the country level and internationally. This supports the existing categorizations on the context scales for the middle region, indicating that most respondents are likely in the "Medium" region of the scale. Overall, the categorization based on this method provides more value for interpreting home early literacy activities with students' reading achievement.

## Discussion

With growing interest in understanding how learning contexts relate to student achievement, many items in large-scale assessment questionnaires are designed to measure a common underlying context construct linked to achievement. For interpretation, respondents

**Table 3:**

*Percent of Students in Each Region of Home Early Literacy Activities Scale Using Two Categorization Methods*

| Country | Percent of Students (Judgment-based Method) | | | Percent of Students (LCA-CS Method) | | | Average Scale Score |
|---|---|---|---|---|---|---|---|
| | High | Medium | Low | High | Medium | Low | |
| Kazakhstan | 66 (0.9) | 34 (0.9) | 0 ~ | 52 (0.9) | 44 (0.8) | 3 (0.5) | 11.3 (0.04) |
| Russian Federation | 64 (1.3) | 35 (1.2) | 1 ~ | 51 (1.3) | 44 (1.0) | 6 (0.7) | 11.3 (0.07) |
| Northern Ireland s | 64 (0.9) | 35 (0.9) | 1 ~ | 54 (1.0) | 42 (1.0) | 4 (0.4) | 11.5 (0.04) |
| Georgia | 59 (1.1) | 40 (1.1) | 1 ~ | 45 (1.0) | 49 (1.0) | 5 (0.6) | 11.0 (0.05) |
| Croatia | 58 (1.1) | 42 (1.1) | 0 ~ | 43 (1.1) | 52 (1.1) | 5 (0.5) | 11.0 (0.05) |
| Malta r | 57 (1.2) | 42 (1.2) | 0 ~ | 44 (1.1) | 50 (1.1) | 6 (0.6) | 11.1 (0.05) |
| Albania | 57 (1.5) | 41 (1.4) | 2 ~ | 44 (1.6) | 47 (1.7) | 9 (1.3) | 10.9 (0.08) |
| Uzbekistan | 57 (1.7) | 43 (1.7) | 0 ~ | 40 (1.6) | 55 (1.4) | 4 (0.5) | 10.8 (0.06) |
| Ireland | 56 (1.1) | 43 (1.0) | 1 ~ | 43 (1.1) | 50 (0.9) | 7 (0.6) | 11.0 (0.05) |
| Kosovo | 55 (1.3) | 44 (1.3) | 1 ~ | 40 (1.3) | 55 (1.2) | 5 (0.6) | 10.8 (0.04) |
| Montenegro | 55 (0.9) | 45 (0.9) | 0 ~ | 41 (0.8) | 54 (0.7) | 5 (0.4) | 10.9 (0.03) |
| North Macedonia | 55 (1.2) | 43 (1.2) | 2 ~ | 43 (1.2) | 51 (1.2) | 6 (1.1) | 10.9 (0.09) |
| Serbia | 54 (1.2) | 46 (1.2) | 0 ~ | 40 (1.3) | 54 (1.0) | 5 (1.0) | 10.8 (0.05) |
| Poland | 53 (0.9) | 47 (1.0) | 0 ~ | 40 (0.9) | 54 (1.0) | 6 (0.5) | 10.8 (0.04) |
| Spain | 52 (0.8) | 47 (0.8) | 1 ~ | 39 (0.8) | 53 (0.8) | 8 (0.4) | 10.7 (0.03) |
| Italy | 52 (0.9) | 47 (0.9) | 1 ~ | 39 (0.8) | 54 (0.8) | 7 (0.4) | 10.7 (0.03) |
| Cyprus | 51 (0.6) | 48 (0.7) | 1 ~ | 39 (0.7) | 53 (0.7) | 9 (0.5) | 10.7 (0.03) |
| Slovak Republic | 49 (1.1) | 49 (1.2) | 2 ~ | 36 (1.0) | 54 (1.2) | 10 (1.6) | 10.5 (0.07) |
| Slovenia | 49 (1.0) | 51 (1.0) | 1 ~ | 37 (0.9) | 55 (0.8) | 8 (0.6) | 10.6 (0.04) |
| Latvia | 48 (1.1) | 51 (1.1) | 1 ~ | 35 (0.9) | 57 (1.0) | 8 (0.5) | 10.5 (0.04) |
| Israel s | 47 (1.0) | 52 (1.0) | 1 ~ | 36 (1.0) | 54 (0.9) | 10 (0.7) | 10.6 (0.04) |
| Hungary r | 47 (1.0) | 52 (1.0) | 1 ~ | 31 (0.9) | 61 (1.0) | 8 (0.6) | 10.5 (0.03) |
| Czech Republic | 46 (0.8) | 54 (0.8) | 0 ~ | 33 (0.8) | 60 (0.8) | 7 (0.4) | 10.5 (0.03) |
| United Arab Emirates s | 42 (0.7) | 56 (0.7) | 2 ~ | 31 (0.6) | 56 (0.5) | 13 (0.4) | 10.3 (0.03) |
| Bulgaria | 41 (1.1) | 50 (1.1) | 9 (1.2) | 30 (1.0) | 50 (1.3) | 20 (1.5) | 9.9 (0.09) |
| France | 41 (0.9) | 57 (0.9) | 2 ~ | 30 (0.9) | 58 (0.9) | 12 (0.6) | 10.2 (0.04) |
| Denmark | 41 (0.9) | 58 (0.9) | 1 ~ | 28 (0.9) | 60 (0.9) | 12 (0.6) | 10.3 (0.04) |
| Germany s | 40 (1.1) | 59 (1.1) | 1 ~ | 27 (1.0) | 64 (1.1) | 9 (0.6) | 10.3 (0.04) |
| Norway (5) | 39 (0.7) | 59 (0.7) | 1 ~ | 28 (0.7) | 61 (0.7) | 11 (0.5) | 10.2 (0.03) |
| Saudi Arabia r | 39 (1.0) | 58 (1.1) | 3 (0.4) | 29 (0.9) | 58 (1.1) | 13 (0.7) | 10.2 (0.05) |
| South Africa r | 38 (0.9) | 58 (0.8) | 4 (0.5) | 28 (0.8) | 56 (0.9) | 16 (0.8) | 10.1 (0.05) |
| Bahrain | 38 (0.7) | 60 (0.7) | 2 ~ | 26 (0.9) | 60 (0.9) | 14 (0.6) | 10.1 (0.03) |
| Sweden s | 38 (1.1) | 61 (1.1) | 1 ~ | 27 (0.9) | 59 (1.0) | 13 (0.9) | 10.2 (0.04) |
| Austria | 37 (0.9) | 61 (0.9) | 1 ~ | 25 (0.8) | 62 (1.1) | 13 (0.8) | 10.1 (0.04) |
| Portugal | 37 (0.9) | 62 (0.9) | 1 ~ | 25 (0.8) | 62 (0.7) | 12 (0.5) | 10.1 (0.03) |
| Azerbaijan | 36 (1.0) | 62 (1.0) | 2 ~ | 23 (0.9) | 63 (1.0) | 14 (0.8) | 10.1 (0.05) |
| Singapore | 35 (0.8) | 62 (0.8) | 4 (0.3) | 26 (0.7) | 54 (0.7) | 19 (0.6) | 10.0 (0.04) |
| Oman | 34 (1.0) | 65 (1.0) | 2 ~ | 21 (0.9) | 67 (1.0) | 11 (0.7) | 10.0 (0.04) |
| Qatar r | 33 (1.0) | 65 (1.0) | 2 ~ | 23 (1.0) | 63 (0.9) | 14 (0.8) | 9.9 (0.04) |
| Finland | 33 (0.7) | 66 (0.7) | 1 ~ | 23 (0.7) | 65 (0.9) | 12 (0.5) | 10.0 (0.02) |
| Turkiye | 31 (1.1) | 57 (1.2) | 13 (1.6) | 22 (1.0) | 52 (1.4) | 27 (1.8) | 9.3 (0.12) |
| Belgium (French) r | 30 (1.0) | 67 (1.0) | 2 ~ | 20 (0.9) | 63 (1.1) | 17 (0.8) | 9.8 (0.04) |
| Brazil | 30 (1.0) | 63 (1.2) | 7 (0.9) | 21 (0.9) | 55 (1.3) | 24 (1.1) | 9.6 (0.06) |
| Jordan | 29 (1.0) | 66 (0.9) | 5 (0.6) | 19 (0.8) | 61 (1.1) | 20 (1.2) | 9.6 (0.06) |
| Belgium (Flemish) | 27 (0.8) | 71 (0.9) | 2 ~ | 17 (0.7) | 62 (0.8) | 21 (0.9) | 9.6 (0.04) |
| Egypt | 27 (1.3) | 67 (1.3) | 7 (0.7) | 17 (1.1) | 60 (1.2) | 23 (1.3) | 9.4 (0.07) |
| Iran, Islamic Rep. of | 24 (1.1) | 71 (1.2) | 5 (0.9) | 15 (0.9) | 61 (1.2) | 24 (1.2) | 9.4 (0.07) |
| Chinese Taipei | 18 (0.5) | 76 (0.6) | 6 (0.4) | 12 (0.4) | 60 (0.7) | 28 (0.7) | 9.1 (0.03) |
| Hong Kong SAR | 16 (0.8) | 81 (0.8) | 3 (0.3) | 10 (0.7) | 66 (0.8) | 24 (0.8) | 9.2 (0.04) |
| Morocco | 13 (0.7) | 67 (1.4) | 19 (1.6) | 8 (0.5) | 49 (1.5) | 44 (1.7) | 8.2 (0.10) |
| Macao SAR | 10 (0.4) | 85 (0.4) | 5 (0.3) | 6 (0.3) | 58 (0.8) | 36 (0.7) | 8.7 (0.02) |
| **International Average** | **42 (0.1)** | **55 (0.1)** | **3 (0.1)** | **31 (0.1)** | **56 (0.1)** | **13 (0.1)** | |
| New Zealand x | 59 (1.1) | 40 (1.1) | 1 ~ | 49 (1.1) | 45 (1.0) | 7 (0.5) | 11.2 (0.05) |
| Netherlands x | 39 (1.3) | 60 (1.4) | 1 ~ | 27 (1.2) | 62 (1.2) | 11 (0.8) | 10.2 (0.05) |

An "r" indicates data are available for at least 70% but less than 85% of the students.
An "s" indicates data are available for at least 50% but less than 70% of the students.
An "x" indicates data are available for at least 40% but less than 50% of the students—interpret with caution.
A tilde (~) indicates insufficient data to report result. A dash (-) indicates comparable data not available.

are classified into high, middle, and low regions utilizing specified cut-points on the context scale. The achievement in each group is then reported. This enables the relationship between achievement and the context to be observed across diverse groups. Conventional methods rely on expert judgments to define cut points based on item content, which works well with balanced response distributions. However, when the item responses are highly skewed across diverse groups or populations, these content-referenced cut-score definitions likely produce regions with few or no respondents, limiting the interpretation of the achievement and context relationship, as illustrated in Table 3.

The proposed LCA-CS method addresses these challenges by leveraging LCA to calculate the posterior probability of class membership for a pre-specified number of classes for each respondent with complete responses. With the assumption that each class is a homogeneous group with a conditional normal ability distribution, the conditional probability approximations of class membership are obtained by a series of calculations, as illustrated in the previous sections. These conditional probabilities of a class membership given a score provide the basis for finding the cut scores on the constructed context scale to apply to all respondents with a valid scale score. As demonstrated by applying the method to the PIRLS 2021 Home Learning Activity data, the proposed LCA-CS method statistically optimized the distribution of students across categories and enhanced the adequacy of categorization. This implies that this data-driven LCA-CS method could serve as an improved approach for identifying cut scores for educational researchers or practitioners, especially when the responses are highly skewed across diverse groups.

Our study aligns with the growing body of literature emphasizing the importance of incorporating statistical modeling techniques into educational assessment to enhance the validity of classification decisions (e.g., Brown, 2007; Templin & Jiao, 2012; Binici & Cuhadar, 2022). While both Brown (2007) and Binici and Cuhadar (2022) focused on the application of LCA-based method to achievement data demonstrating its utility as an empirical, data-driven alternative to judgement-based methods for classifying examinees, our research extends the application of LCA-based method to contextual data. In this domain, where response distributions are often skewed across diverse groups, LCA-based classifications can improve the adequacy of categorization. Furthermore, our findings resonate with those of Binici and Cuhadar (2022), who demonstrated that LCA-based methods can validate performance standards derived from traditional judgment-based approaches. Similarly, in the context of our study, the LCA-based method proved effective for validating existing judgement-based categorizations on the context scales.

In conclusion, the LCA-CS method offers a promising, statistically sound alternative for defining cut scores on context scales in large-scale assessments. By addressing the limitations of traditional methods and optimizing the distribution of respondents across categories, this approach provides meaningful insights into the relationship between learning contexts and achievement. The LCA-CS method, as introduced in this study, utilized scales derived from a Rasch model with a pre-specified number of classes provided by analytic goals. When this required number of classes is unavailable, the LCA method can be used to determine the optimal number of classes based on model fit statistics and practical needs. This study introduced the LCA-CS method and demonstrated its implementation with real data from a large-scale assessment. Future studies should focus on developing diagnostics to evaluate the effectiveness of the LCA-CS method compared to judgment-based cut points. In addition, future research could extend this approach to scales based on more general IRT models, such as the Generalized Partial Credit Model, using similar procedures.

**References**

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika* 42, 69–81. https://link.springer.com/content/pdf/10.1007/BF02293746.pdf

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Binici, S., & Cuhadar, I. (2022). Validating performance standards via latent class analysis. *Journal of Educational Measurement*, 59(4), 502-516.

Brown, R. S. (2007). Using latent class analysis to set academic performance standards. *Educational Assessment*, 12(3-4), 283-301.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, 30(2), 93–106. https://doi.org/10.1111/j.1745-3984.1993.tb01068.x

Cizek, G. J. (2005). EMW. *Defending Standardized Testing*, 23.

Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations*. Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications Ltd.

Clogg, C. C. (1979). Some latent structure models for the analysis of Likert-type data. *Social Science Research*, 8(4), 287-301.

Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology*, 43(2), 171-192.

Croon, M. (2002). Ordering the classes. *Applied latent class analysis*, 137-162.

Dayton, C. M., & Macready, G. B. (1976). A probabilistic model for validation of behavioral hierarchies. *Psychometrika*, 41(2), 189–204. https://doi.org/10.1007/BF02291838

Dayton, C. M., & Macready, G. B. (2006). 13 Latent Class Analysis in Psychometrics. *Handbook of statistics*, 26, 421-446.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, 87(418), 476-486.

Heinen, T. (1993). *Discrete Latent Variable Models*. Tilburg: University Press.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences. Sage Publications*, Inc.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (pp. 485–514). New York: Macmillan.

Jiao, H., Lissitz, R. W., Macready, G., Wang, S. & Liang, S. (2011). Exploring levels of performance using the mixture Rasch model for standard setting. *Psychological Test and Assessment Modeling*, 53(4), 499-522.

Lanza, S. T., Dziak, J. J., Huang, L., Wagner, A. T., & Collins, L. M. (2015). *Proc LCA & Proc LTA users' guide* (Version 1.3.2). University Park: The Methodology Center, Penn State. Available from methodology.psu.edu.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. New York, NY: Houghton Mifflin

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education*, 2(2), 121–141.

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, 47(2), 149-174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In: W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*. Berlin: Springer.

Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational Statistics*, 8(4), 271-288.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2013). The bookmark procedure: Psychological perspectives. In *Setting Performance Standards* (pp. 263-296). Routledge.

Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., & Wry, E. (2023). *PIRLS 2021 International Results in Reading*. Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb5342

Peabody, M. R., Muckle, T. J., & Meng, Y. (2023). Applying a Mixture Rasch Model-Based Approach to Standard Setting. *Educational Measurement: Issues and Practice*, *42*(3), 5-12.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271-282.

Rost, J., & von Davier, M. (1995). Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), Rasch models – Foundations, Recent Developments, and Applications (pp. 371–379). Springer. https://doi.org/10.1007/978-1-4612-4230-7_20

Smit, J. A., Kelderman, H., & van der Flier, H. (2003). Latent trait latent class analysis of an Eysenck Personality Questionnaire. *Methods of Psychological Research Online*, 8(3), 23-50.

Templin, J., & Jiao, H. (2012). Applying model-based approaches to identify performance categories. In *Setting performance standards* (pp. 379-397). Routledge.

von Davier, M., & Lee, Y.-S. (2019). Introduction: From latent classes to cognitive diagnostic models. In Handbook of Diagnostic Classification Models (pp. 1-17). Springer. https://doi.org/10.1007/978-3-030-05584-4_1

Yamamoto, K. (1987). *A model that combines IRT and latent class models*. University of Illinois at Urbana-Champaign.

Yin, L., & Reynolds, K. A. (2023). Creating and interpreting the PIRLS 2021 context questionnaire scales. In M. von Davier, I. V. S. Mullis, B. Fishbein, & P. Foy (Eds.), *Methods and Procedures: PIRLS 2021 Technical Report* (pp. 15.1-15.161). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb6994

iejee

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

# Latent Profile Analysis: Comparison of Achievement versus Ability-Derived Subgroups of Mathematical Skills

Onur Demirkaya[a,*], Sharon Frey[b], Sid Sharairi[c], JongPil Kim[d]

Received : 14 December 2024
Revised : 2 February 2025
Accepted : 2 March 2025
DOI : 10.26822/iejee.2025.379

[a] **Corresponding Author:** Onur Demirkaya, Riverside Insights, Research and Measurement Services, Illinois, USA.
E-mail: onur.demirkaya@riversideinsights.com
ORCID: https://orcid.org/0000-0002-3985-6485

[b] Sharon Frey, Riverside Insights, Research and Measurement Services, Illinois, USA.
E-mail: sharon.frey@riversideinsights.com
ORCID: https://orcid.org/0009-0002-2597-674X

[c] Sid Sharairi, Riverside Insights, Research and Measurement Services, Illinois, USA.
E-mail: sid.sharairi@riversideinsights.com
ORCID: https://orcid.org/0009-0005-6045-0403

[d] JongPil Kim, Riverside Insights, Research and Measurement Services, Illinois, USA.
E-mail: jp.kim@riversideinsights.com

## Abstract

This study compares latent profiles derived from student subgroups of varying levels of mathematical skills defined by achievement and ability assessment scores. Achievement and ability cut scores for identifying students at both ends of the mathematics spectrum were applied and the resulting latent profiles within each condition were compared. The research utilized latent profile analysis to identify student profiles with achievement scores from the Iowa Assessments and ability scores from CogAT. The participants consisted of 50,998 second-grade students in a Southeastern state. The finding revealed varying demographics and patterns of ability and achievement for each condition, underscoring the need to acknowledge students with diverse learning styles and the distinct dynamics between achievement and ability scores to use for identifying students who may benefit from tailored educational programs.

**Keywords:**

Mathematical Skills, Ability Assessment, Achievement Assessment, Latent Profile Analysis, CogAT

## Introduction

The COVID-19 pandemic significantly disrupted teaching and learning processes, leading to notable declines in student achievement across grade levels. Numerous reports have examined the pandemic's impact, consistently highlighting that mathematics achievement suffered more than reading (Curriculum Associates, 2020; Kuhfeld et al., 2020; Renaissance Learning, 2021). Even prior to the pandemic, academic performance in the United States revealed concerning trends, with 30% of Grade 12 students performing below the basic level in reading and 40% below the basic level in mathematics (National Center for Educational Statistics, 2019). Mathematics, especially, poses challenges for many students and often serves as a gatekeeper to higher education and employment opportunities in technology-driven fields (Moses & Cobb, 2001). The cumulative nature of mathematical learning, where advanced concepts build on foundational skills, further exacerbates difficulties for students who fall behind, making it challenging for them to catch up with their peers (Green et al., 2017).

Given these challenges, understanding how to enhance academic achievement, particularly in mathematics and reading, is a pressing concern for parents, educators, and policymakers (Younger, et al., 2024). Developing targeted strategies to support skill acquisition in these areas is essential, as they form the foundation for broader educational and professional success. Understanding and addressing these issues is important to improve outcomes and ensure equitable opportunities for all students.

In many educational systems, students are traditionally grouped based on cognitive abilities, achievement scores, and other measures to provide more targeted instruction to students with shared strengths or weaknesses. These categories often include students identified as gifted and talented or those participating in individual or intervention education programs. While such groups are more homogenous in terms of selection criteria, studies show that diverse profiles often arise due to various factors reflecting a range of educational, cognitive and social influences (e.g., Mahatmya et al., 2023; Mammadov et al., 2016; Ziernwald e al, 2022) . For instance, some students may excel in specific areas (e.g., math, verbal reasoning) but not necessarily across all domains. "Twice-exceptional" students – those who are both gifted and have learning disabilities – may show discrepancies between achievement and ability scores (Moon & Reis, 2004). Socioeconomic background also plays a role, for example, with high-SES students often benefiting from more exposure to advanced learning resources, resulting in higher achievement scores, while low-SES students may underperform despite having high ability.

Another source of diversity with these groups arises from the tools used to identify students, such as achievement and ability tests along with other measures. Therefore, it is important to distinguish between achievement and ability, as these constructs, while related, assess different aspects of student performance. Achievement typically refers to the knowledge and skills a student has acquired through learning and education, often reflected through test scores and grades (Soares, et al., 2015). In contrast, ability–sometimes referred to as fluid intelligence (Cattell, 1963, 1987)–is typically measured by tests of inductive and deductive reasoning, assessing a student's potential to think critically, solve problems, draw inferences, identify relationships, and transform information in a significant way (Nickerson, 2011). That is, the ability reflects potential, whereas achievement represents the realization or execution of that potential (Schneider, 2013). Understanding the differences between these two constructs is essential for accurately identifying students' needs, as a high-achieving student may not necessarily possess the highest levels of innate ability, and vice versa.

The association between ability and academic achievement is well-established. A large body of research has demonstrated a significant correlation between ability and achievement, ranging from .50 to .70 (Soares, et al., 2015). Variable-centered approaches (e.g., analytic approaches that examine associations among variables; Laursen & Hoff, 2006), such as an ordinary least squares regression, may offer a limited perspective of student performance, potentially obscuring significant subgroups with unique achievement and ability performance patterns because they focus on inter-individual differences instead of intra-individual differences (Litkowski, et al., 2020). In contrast, latent profile analysis (LPA), a person-centered approach, identifies groups of individuals who share certain characteristics (Laursen & Hoff, 2006). By clustering students into latent profiles that reflect shared characteristics across achievement and ability metrics, LPA provides a more nuanced understanding of student diversity and performance.

The existing literature includes studies examining latent profiles of critical thinking and science achievement (Hwang et al., 2023), as well as cognitive profiles based on executive functioning to predict academic performance in reading and mathematics (Carriedo, et al., 2024; Younger, et al., 2024; Litkowski, et al., 2020), and exploration of latent profiles of mathematics achievement, numerosity, and math anxiety in twins (Hart et al., 2016). Additionally, research has explored unique profiles of high-ability and underrepresented students' subject-specific psychological strengths (Mahatmya et al., 2023) and has emphasized the role of LPA in understanding personality profiles of high ability students *L-Ach* (Mammadov et al., 2016). Furthermore, Zierwald et al. (2022) utilized the LPA to differentiate high-achieving subgroups based on different mathematic achievement indicators and the motivational-affective characteristics. Despite these contributions, to our knowledge, thus far, no study has explicitly addressed the heterogeneity in students' performance across both achievement and all components of reasoning ability scores, particularly within the context of high- and low-performing groups.

Therefore, this study aims to explore how high- and low-performing groups, as defined by standardized achievement and ability test scores, differ in their latent profiles derived from standardized achievement (Mathematics and Reading) and ability (Verbal, Quantitative and Nonverbal) tests scores. Specifically, it seeks to answer four major research questions:

1. Do low-achieving and low-ability groups, as defined by achievement and ability test scores, have configural differences (number and shape of profiles) in the latent profiles derived?

2. Do high-achieving and high-ability groups, as defined by achievement and ability test scores,

have configural differences in the latent profiles derived?

3. What are the demographics of students within each of the latent profiles?

4. How do the patterns of test and skill level performances compare across student profiles?

Understanding these profiles has significant implications for educational practitioners. For instance, recognizing that students may differ significantly in terms of learning preferences, strengths, or areas of struggle can inform the design of differentiated instruction, more targeted interventions or support mechanisms tailored to address each subgroup's specific needs. By focusing on both ends of the achievement and ability spectrum, this study offers comprehensive insights into how these student groups differ not just on performance measures but also in their latent academic profiles, potentially guiding future educational policies and practices.

**Method**

Participants. This study utilized one year of data from one large, diverse school district in the Southeast United States. The data contained 55,482 Grade 2 students who tested with both an achievement and an ability assessment in October of 2022. After excluding individuals who failed to complete the test, encountered testing irregularities, or lacked scores in any of the Iowa Assessments subjects or any of the Cognitive Abilities Test (CogAT) batteries, the remaining 50,998 (49.8% female) test takers were considered in this study.

The demographics in the study samples were as follows: 64% White, 35.3% Black, 12.7% Hispanic, 3.3% Asian, 1% Pacific Islander, and 3.3% students who identified as American Indian or Alaskan Native. Coding was based on information provided by the district for the CogAT. For the race/ethnicity data fields, students were allowed the option to mark all that apply; therefore, the sum of the percentages may exceed 100%. The demographics and summary statistics of the conditions investigated are provided in the data analysis section.

The second-grade data were selected as this grade provides math instruction that involves a diverse range of foundational skills (see Table A1 in the appendix) and most educational systems administer the CogAT for their gifted/talented screening at this grade level. Institutional Review Board (IRB) approval from [blinded] was not required, as the study involved only secondary data analysis using non-identifiable data elements. However, the researchers did not obtain permission from the school district to make the data publicly accessible. Also, neither student nor district-level information is publicized.

Measures. Data from the following measures were collected as a part of the district's planned assessment schedule. De-identified data from these assessments, along with demographic information were provided for this study.

The Iowa Assessments (Dunbar & Welch, 2015). The achievement test was developed with multiple test levels spanning Grades K to 12 that measure knowledge of subject areas that students are expected to have learned at school (e.g., Reading and Mathematics). The content coverage reflects extensive research by an experienced development team using established professional content standards listed in Table A2 (Riverside Insights, 2012). See Table A1 for the skill domains reported for the test level administered for this study. Students' data from Level 7 of the Iowa Assessments Form G Core Battery: Reading (Part 1—Picture Stories and Sentences and Part 2—Stories) and Mathematics (Part 1 and Part 2) were used in this study. These tests vary in length from 35 to 41 questions, and although the tests are untimed, the estimated time for a student to respond to both parts of a test ranges from 45 to 50 minutes. Except for the Reading test, questions are presented orally. To obtain a Reading score and a Mathematics score, both parts of each of the tests must be administered.

The CogAT (Lohman & Lakin, 2017). The cognitive reasoning ability test was developed to span Grades K to 12 for students aged 4 years 11 months to 21 years 7 months and has two alternate test forms designed to be parallel in test structure and item difficulty. The test assesses inductive and deductive reasoning, classified as fluid-analytic abilities (Cattell, 1963; 1987), in three domain areas—nonverbal/figural, verbal, and quantitative reasoning. These abilities are closely related to an individual's success in school and the test results may be used to help plan adaptable instruction. The data used in this study is from the Level 8 tests of Form 8. For this level, tests vary in length from 14 to 18 questions, and although all the tests are untimed, the estimated time for a student to respond to each test ranges from 11 to 15 minutes.

**Data Analysis**

Two conditions were established to classify students: those scoring in either the lower end (L) or upper end (U) of the score distribution, as determined by norm-referenced scores. The classification was based on the National Percentile Rank (NPR) for either the mathematics test of the Iowa Assessments (mathematics achievement) or the quantitative reasoning battery of the CogAT (quantitative reasoning ability). CogAT provides two types of percentile rank scores: age-based and grade-based. For this study, we utilized the age-based percentile rank. Within each condition, students were identified using achievement (Ach) and ability (Abl) test-based cut scores

corresponding to the 23rd and 77th percentile ranks (Jesson, 2018) for the L and U conditions, respectively. For instance, examinees whose national percentile ranks for the Iowa Mathematics test are lower than or equal to 23 composed the lower achievement group (*L-Ach*), and examinees with age-based national percentile ranks higher than or equal to 77 for the CogAT Quantitative Battery composed the upper ability group (*U-Abl*). Figure 1 displays the subgroups created based on these thresholds.

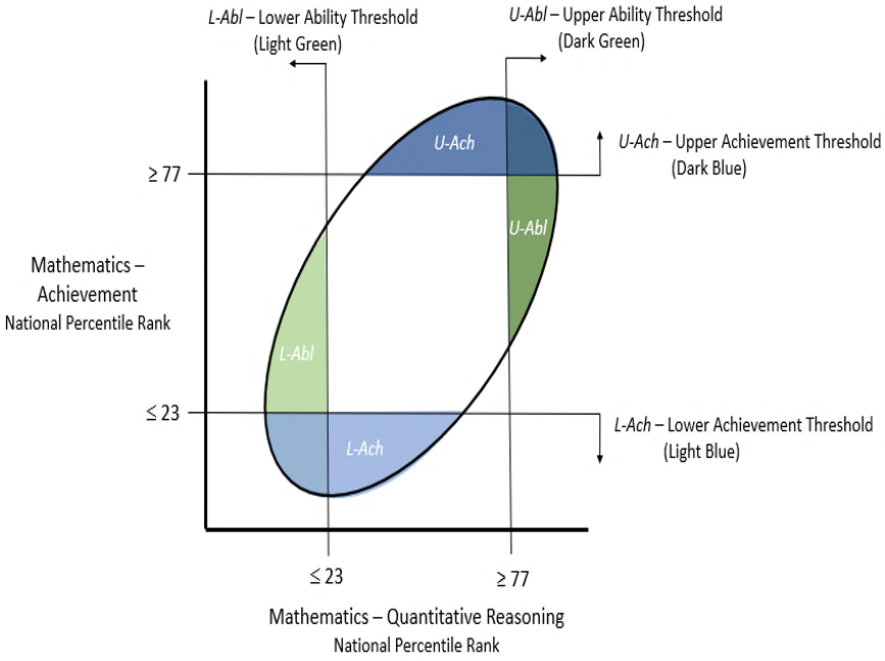
These cut-off scores were selected because they align with the percentile rank thresholds used to define below-average (stanine scores of 1 through 3) and above-average (stanine scores of 7 through 9) performance on both the Iowa and CogAT (Lohman, 2013) assessments. The use of these stanine-based thresholds is particularly relevant because the differentiated instruction reports and profile scores provided by the CogAT assessments are also based on stanine scores (Lohman, 2013). Consequently, these scores are familiar to instructors and have been widely utilized to guide tailored instructional practices.

The demographics of subgroups are provided in Table 1. The achievement-based selection provided the largest sample size in the lower condition while the ability-based criteria selected the largest sample size in the upper condition. Female (52.3%) and black (50.2%) students slightly dominated the *L-Ach* group whereas the *L-Abl* group was slightly dominated by male (52.6%) and black (52.9%) students. Male and white students, on the other hand, dominated both *U-Ach* (60.9%; 87.3%, respectively) and *U-Abl* (56.6%; 80.2%, respectively) groups. In the upper condition, ability-based selection increased the representation of both female and underrepresented groups (Black and Hispanic) compared to the achievement-based selection.

The rescaling of variables before conducting latent profile analysis is a widely common methodological application to ensure interpretable latent profiles (e.g., Carriedo et al., 2024; Spurk et al., 2020). Therefore, the *Iowa Assessments* scale scores (Mathematics and Reading) were rescaled to be on the same scale as the *CogAT* ability normative scale scale ($\bar{x}$ = 100, *SD* = 16). The descriptive statistics of rescaled scores of subgroups *(L-Ach, L-Abl, U-Ach, U-Abl)* are presented in Table 2 to provide an overview of the performance of subgroups on each test. The achievement-based subgroups *(L-Ach & U-Ach)* had higher average test scores than the ability-based subgroups in their specific conditions. In the lower condition, the largest performance differences were on the ability tests whereas the largest performance gaps between the subgroups in the upper condition were on the achievement tests.

To address the research questions, latent profile analyses were conducted using the tidyLPA package (Rosenberg et al, 2019) in R (R Core Team, 2023) for all four subgroups of students (*L-Ach, L-Abl, U-Ach, U-Abl*). Iowa achievement test scores (Iowa Mathematics and Iowa Reading) and *CogAT* ability test scores (Verbal, Quantitative, and Nonverbal Reasoning) were employed to construct student profiles. *LPA* was used as an exploratory-driven approach, and a variety of

**Figure 1.**
*Ability/achievement subgroups based on the thresholds.*



models were investigated to determine the optimum number of profiles. This exploratory-driven approach is appropriate where there is no strong theory to suggest or predict the number of classes or profiles that will result from the underlying variables (Hwang et al., 2023). As with other latent variable models, the model fit indices provided in LPA enable different models to be compared and informed decisions to be made regarding the number of underlying classes which is most congruent with the data (Marsh et al., 2009).

An analytic hierarchy process (Akogul & Erisoglu, 2017), based on the Akaike Information Criterion (AIC, Akaike, 1974), Approximate Weight of Evidence (AWE, Banfield & Raftery, 1993), the Bayesian Information Criterion (BIC, Schwarz, 1978), Classification Likelihood Criterion (CLC, Biernacki & Govaert, 1997), and Kullback Information Criterion (KIC, Cavanaugh, 1999), were examined to determine the optimal number of latent profiles for each set of students. For the model fit indices, models with lower values indicate better fit. In addition to relying on model fit indices, the bootstrap likelihood ratio test (BLRT; McLachlan & Peel, 2000) was utilized to assess model adequacy. A statistically significant BLRT result indicates rejection of the null hypothesis of k profiles in favor of a model with k+1 profiles. Other considerations in selecting the optimal model included profile sizes (Lubke & Neale, 2006) and the interpretability of the profiles (Marsh et al., 2009). After identifying the final model, the descriptive statistics and prevalence of each profile were summarized and examined. The latent profiles resulting from the achievement versus ability test-based cut scores were compared for both conditions

**Table 1.**

*Demographic Distributions of the Matched Datasets by Condition and Subgroup.*

| Condition | Subgroup | *N* | Female | Male | American Indian | Asian | Black | Hispanic | Pacific Islander | White | Other |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Lower | *L-Ach (Math NPR ≤ 23)* | 22288 | 52.3% | 47.6% | 4.5% | 2.1% | 50.2% | 17.4% | 1.2% | 49.0% | 1.1% |
| | *L-Abl (Quant NPR ≤ 23)* | 8650 | 47.2% | 52.6% | 4.1% | 1.2% | 52.9% | 14.8% | 1.1% | 46.8% | 1.3% |
| Upper | *U-Ach (Math NPR ≥ 77)* | 5673 | 39.1% | 60.9% | 1.4% | 6.7% | 10.1% | 5.0% | 0.5% | 87.3% | 0.9% |
| | *U-Abl (Quant NPR ≥ 77)* | 12353 | 43.4% | 56.6% | 2.2% | 6.6% | 16.8% | 9.2% | 0.7% | 80.2% | 1.0% |

**Table 2.**

*Descriptive Statistics for the Matched Datasets by Condition and Subgroup.*

| | | | Achievement | | | | | | | | Ability | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Mathematics | | Reading | | Verbal | | Quantitative | | Nonverbal | |
| Sample | Condition | Subgroup | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Ability/ Achievement Matched Sample | Lower | *L-Ach* | 85.3 | 8.7 | 90.4 | 11.9 | 87.7 | 11.8 | 92.5 | 12.3 | 88.0 | 11.4 |
| | | *L-Abl* | 83.9 | 11.5 | 87.5 | 11.5 | 81.2 | 11.1 | 79.2 | 7.6 | 80.6 | 9.6 |
| | Upper | *U-Ach* | 127.1 | 6.1 | 117.8 | 13.1 | 113.1 | 10.5 | 118.3 | 10.0 | 116.7 | 13.2 |
| | | *U-Abl* | 114.8 | 12.2 | 111.4 | 14.7 | 109.3 | 10.7 | 119.3 | 6.3 | 113.3 | 12.6 |
| | Overall | Total Group | 100.0 | 16.0 | 100.0 | 16.0 | 96.9 | 14.1 | 102.0 | 14.3 | 97.6 | 15.4 |

Note: The Iowa Assessments scale scores (Mathematics and Reading) were rescaled to be on the same scale as the CogAT ability normative scale ($\bar{x}$ = 100, *SD* = 16). The total group is comprised of all examinees (*N=50998*) in the matched sample.

**Table 3.**

*Model Fit Statistics for Models for Each Condition and Subgroup.*

| Condition | Subgroup | Model | LL | AIC | BIC | Entropy | n-min% | BLRT |
|---|---|---|---|---|---|---|---|---|
| Lower | *L-Ach* | 1 | -409398.89 | 818837.79 | 818998.03 | 1.00 | 100.00% | n/a |
| | | 2 | -406727.98 | 813537.95 | 813866.43 | 0.70 | 32.42% | p<.01 |
| | | 3 | -405812.49 | 811748.98 | 812245.71 | 0.61 | 29.69% | p<.01 |
| | | 4 | -405604.68 | 811375.37 | 812040.35 | 0.54 | 6.42% | p<.01 |
| | | **5** | **-405430.73** | **811069.46** | **811902.69** | **0.49** | **16.63%** | **p<.01** |
| | | 6 | n/a | n/a | n/a | n/a | n/a | n/a |
| | *L-Abl* | 1 | -157485.99 | 315011.98 | 315153.28 | 1.00 | 100.00% | n/a |
| | | 2 | -155936.00 | 311953.99 | 312243.67 | 0.55 | 46.80% | p<.01 |
| | | 3 | -155589.40 | 311302.81 | 311740.86 | 0.53 | 18.55% | p<.01 |
| | | 4 | -155245.08 | 310656.16 | 311242.58 | 0.56 | 20.18% | p<.01 |
| | | **5** | **-155112.84** | **310433.69** | **311168.48** | **0.53** | **9.78%** | **p<.01** |
| | | 6 | n/a | n/a | n/a | n/a | n/a | n/a |
| Upper | *U-Ach* | 1 | -103674.33 | 207388.66 | 207521.53 | 1.00 | 100.00% | n/a |
| | | 2 | -102729.64 | 205541.28 | 205813.66 | 0.50 | 30.88% | p<.01 |
| | | **3** | **-102276.65** | **204677.29** | **205089.19** | **0.66** | **8.27%** | **p<.01** |
| | | 4 | n/a | n/a | n/a | n/a | n/a | n/a |
| | | 5 | n/a | n/a | n/a | n/a | n/a | n/a |
| | | 6 | n/a | n/a | n/a | n/a | n/a | n/a |
| | *U-Abl* | 1 | -227326.56 | 454693.13 | 454841.56 | 1.00 | 100.00% | n/a |
| | | 2 | -225648.74 | 451379.49 | 451683.78 | 0.54 | 40.33% | p<.01 |
| | | 3 | -224991.37 | 450106.74 | 450566.88 | 0.58 | 25.58% | p<.01 |
| | | **4** | **-224642.41** | **449450.83** | **450066.82** | **0.65** | **11.31%** | **p<.01** |
| | | 5 | -224279.65 | 448767.30 | 449539.15 | 0.63 | 11.06% | p<.01 |
| | | 6 | -224216.03 | 448682.06 | 449609.76 | 0.61 | 10.53% | p<.01 |

Note: Bolded is the selected model. LL = Log-likelihood; AIC = Akaike information criteria; BIC = Bayesian information criteria; n-min% = the profile with the smallest percentage of individuals assigned to it; BLRT = The Bootstrap Likelihood Ratio Test; n/a = used to represent nonconvergence or not applicable conditions.

(Lower: *L-Ach* vs. *L-Abl* and Upper: *U-Ach* vs. *U-Abl*). To address the third research question, the percentage distribution of individuals within each profile across demographic categories (e.g., gender and ethnicity) was analyzed. For the final research question, Reading and Mathematics skill scores were summarized across profiles and conditions to compare their patterns to both that of the national averages and within each condition.

**Results**

A series of LPA models with various constraints (EEI: Equal variances and zero covariances; VVI: Varying variances and zero covariances; EEE: Equal variances and equal covariances; VVV: Varying variances and varying covariances) and up to six profile solutions were run to examine and determine the number of latent profiles for each subgroup. Among all models, solutions with the VVV model provided the best model fit statistics than the others. That is expected since the VVV model is less parsimonious than all the other models yet has the potential to allow for understanding many aspects of the variables that are used to estimate the profiles (Rosenberg et al, 2019). Therefore, fit indices for each solution with only the VVV model are reported in Table 3.

The analytic hierarchy process suggested a five-profile solution for *L-Ach, L-Abl and U-Abl* subgroups but three profiles for the *U-Ach* group. Four, five, and six-profile solutions with the VVV model did not converge for *U-Ach* whereas a six-profile solution did not converge for the *L-Ach*, and *L-Abl*. Even though the fit indices supported a five-profile solution over a four-profile solution for the *U-Abl* subgroup (*BIC* = 449539.15; entropy = 0.63; *BLRT* = 776.60; $p < 0.01$), we determined that the fifth profile had already been represented by another profile with a very slight difference in means at three points (Mathematics, Verbal, and Quantitative). Therefore, the fifth profile did not add meaningful and important information about the heterogeneity in this subgroup. Table 4 provides the mean and standard deviations, as well as the corresponding proportions for each of the latent profiles across the conditions.

Figures 2 & 3; and 4 & 5 visually depict the profiles of the subgroups at the lower and upper conditions, respectively. As is typical in LPA, the naming of profiles is informed by the shape of the profiles. After a thorough examination of Figures 2, 3, and Table 4, we decided that the profile distinction was based on both the general relative performance across the achievement and ability tests and the relative performance between the achievement tests for the *L-Ach* group. These labels are (a) high performance (High), (b) medium performance (Medium), (c) medium performance with Reading strength (Medium-RS), (d) low performance (Low), and (e) low performance with Math weakness (Low-MW). For the *L-Abl* group, the achievement performances were generally higher than the ability performances within profiles (Ach > Abl). Therefore, the distinction was based on the relative performance comparison between the achievement and ability tests for this subgroup. These profile labels are (a) high achievement-high ability

**Table 4.**

*Descriptive Statistics for Achievement and Ability Measures with Sample Sizes Across Latent Profiles and Subgroups.*

| | | | | Achievement | | | | | | | | Ability | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sample Size | | Reading | | Mathematics | | Verbal | | Quantitative | | Nonverbal | |
| Subgroup | Profile | *N* | *%* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| *L_Ach* | High | 6020 | 27.0% | 96.4 | 12.4 | 94.9 | 1.8 | 95.0 | 9.5 | 99.7 | 9.3 | 94.2 | 10.7 |
| | Medium | 4316 | 19.4% | 89.4 | 8.6 | 88.6 | 4.2 | 88.9 | 9.5 | 95.5 | 9.7 | 88.0 | 8.9 |
| | Medium-RS | 3968 | 17.8% | 94.8 | 14.4 | 84.8 | 5.1 | 91.9 | 9.2 | 96.0 | 8.6 | 90.2 | 9.8 |
| | Low | 4278 | 19.2% | 85.7 | 9.2 | 81.7 | 7.1 | 81.3 | 11.9 | 84.8 | 13.4 | 84.1 | 13.5 |
| | Low-MW | 3706 | 16.6% | 84.8 | 8.8 | 73.1 | 5.8 | 80.4 | 11.2 | 85.3 | 11.4 | 82.2 | 7.8 |
| *L_Abl* | HAHA | 1610 | 18.6% | 91.9 | 12.7 | 92.2 | 10.8 | 85.8 | 9.4 | 83.1 | 2.4 | 83.2 | 10.9 |
| | High | 2035 | 23.5% | 89.4 | 11.7 | 87.4 | 11.1 | 86.0 | 9.5 | 86.8 | 1.1 | 84.6 | 7.6 |
| | HAMLA | 846 | 9.8% | 91.8 | 14.9 | 88.2 | 10.7 | 82.6 | 9.9 | 74.1 | 6.3 | 79.6 | 10.2 |
| | Medium | 2445 | 28.3% | 83.9 | 8.6 | 79.5 | 8.6 | 79.6 | 11.1 | 80.4 | 3.8 | 81.3 | 8.9 |
| | Low | 1714 | 19.8% | 83.2 | 6.8 | 75.4 | 7.4 | 73.8 | 10.3 | 70.5 | 7.8 | 74.3 | 9.2 |
| *U_Ach* | High | 469 | 8.3% | 124.2 | 10.9 | 138.7 | 6.5 | 119.3 | 10.6 | 124.4 | 8.8 | 122.8 | 12.4 |
| | Medium | 2383 | 42.0% | 118.5 | 12.7 | 128.2 | 3.4 | 113.9 | 10.2 | 119.0 | 9.7 | 118.1 | 13.1 |
| | Low | 2821 | 49.7% | 115.0 | 13.2 | 122.3 | 1.5 | 110.3 | 9.8 | 115.5 | 9.8 | 113.1 | 12.6 |
| *U_Abl* | High-RS | 1397 | 11.3% | 132.8 | 4.6 | 120.5 | 9.9 | 114.4 | 10.0 | 120.0 | 4.6 | 117.2 | 11.4 |
| | High-QS | 3475 | 28.1% | 113.0 | 13.4 | 120.0 | 11.6 | 112.6 | 10.9 | 125.4 | 6.1 | 119.3 | 12.8 |
| | Medium | 5225 | 42.3% | 107.3 | 12.8 | 111.8 | 11.3 | 107.1 | 9.6 | 116.8 | 2.6 | 110.2 | 11.2 |
| | Low | 2256 | 18.3% | 105.3 | 13.8 | 108.5 | 11.3 | 104.8 | 9.8 | 112.6 | 0.6 | 106.3 | 9.9 |

Note: The Iowa Assessments scale scores (Mathematics and Reading) were rescaled to be on the same scale as the CogAT ability normative scale ( $\bar{x}$ = 100, SD = 16). Medium-RS = Medium Performance with Reading Strength; Low-MW = Low Performance with Math Weakness; HAHA = High Achievement High Ability (Ach>Abl); HAMLA = High Achievement Medium/Low Ability (Ach>Abl); High-QS = High Performance with Quantitative Strength; High-RS = High Performance with Reading Strength.

**Figure 2.**
*Profiles of Low Achievement (L-Ach) Subgroup.*

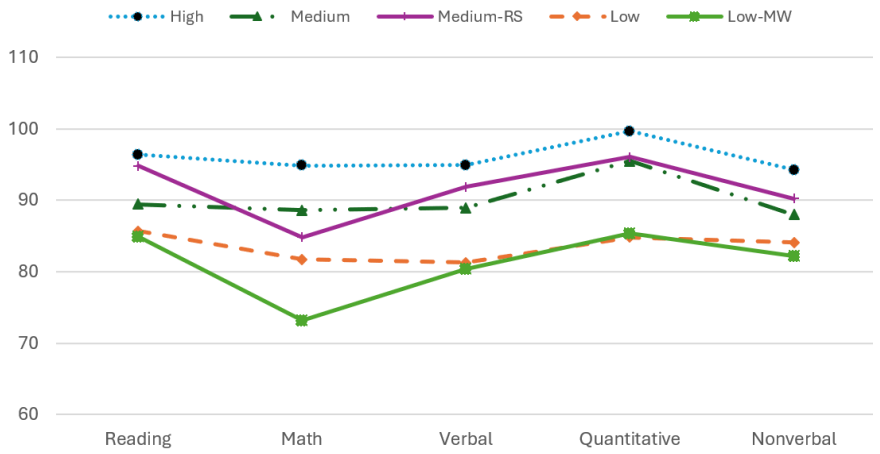
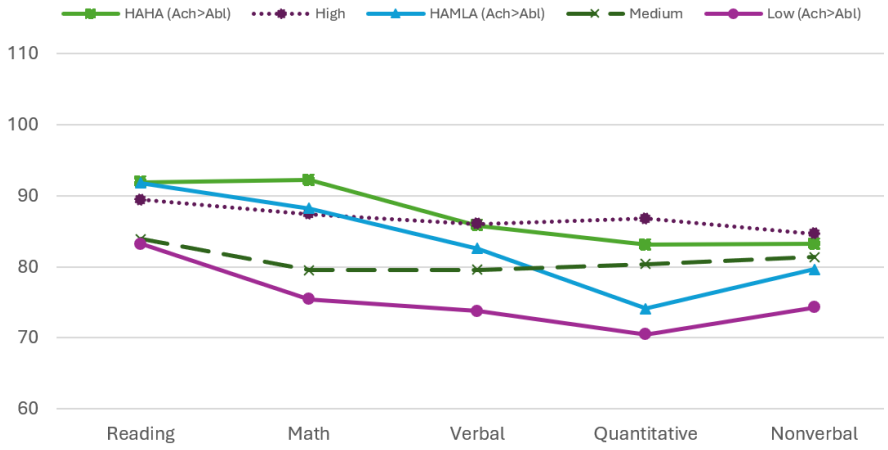



**Figure 3.**
*Profiles of Low Ability (L-Abl) Subgroup.*




(HAHA [Ach > Abl]), (b) high achievement-medium/low ability (HAMLA [Ach > Abl]), (c) high performance (High), (d) medium performance (Medium) and (e) low performance (Low [Ach > Abl]).

Naming the profiles of each subgroup for the upper condition was more straightforward than naming the lower condition. After reviewing Figures 4 and 5, the three profiles identified for the *U-Ach* include (a) a high-performance group (High), (b) a medium-performance group (Medium), and (c) a low-performance group (Low) whereas, for the *U-Abl*, the four profiles identified include (a) a high performance with Reading strength group (High-RS), (b) a high performance with Quantitative strength group (High-QS), (c) a medium-performance group (Medium), and (d) a low-performance group (Low).

Subsequently, the detailed findings were discussed in alignment with the research questions outlined in the introduction.

The analysis of low-achieving and low-ability groups to determine potential configural differences (e.g., number and shape of the profiles) revealed that the number of identified profiles remained stable at five, although the patterns within these profiles demonstrated variation. This indicates that the underlying characteristics and interactions between performance metrics differ depending on whether the group is defined by achievement outcomes or inherent ability measures at the lower percentile examinees.

Among the low-achieving group, students displayed relatively lower performance in mathematics compared to their quantitative reasoning abilities, particularly within the Medium-RS and Low-MW profiles. This discrepancy indicates that these profiles may represent students who are underperforming in mathematics relative to their potential in quantitative reasoning. This highlights potential unmet educational needs or contextual barriers affecting mathematics achievement for students in this group. This discrepancy underscores the importance of tailored interventions that bridge the gap between potential and performance.

In the low-ability group, profile patterns were generally consistent across domains; however, notable dips were observed in Quantitative performance for

**Figure 4.**
*Profiles of Upper Achievement (U-Ach) Subgroup.*

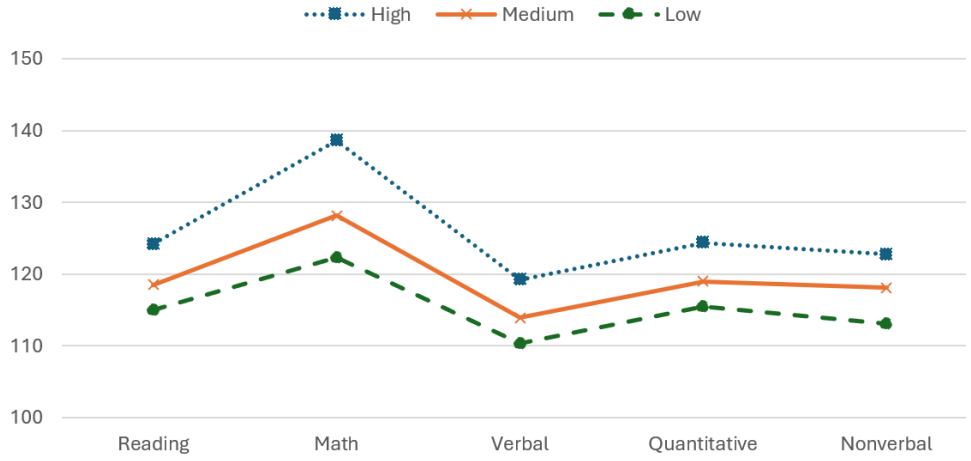



**Figure 5.**
*Profiles of Upper Ability (U-Abl) Subgroup.*

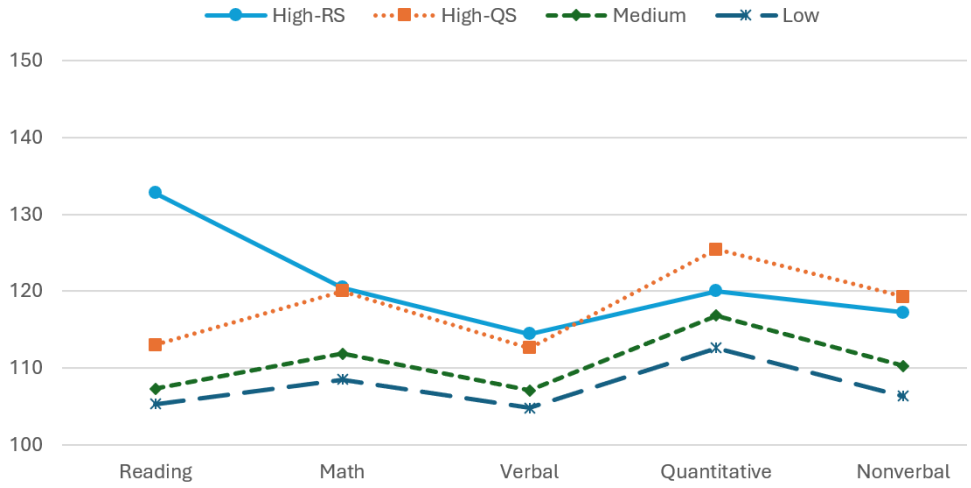



**Table 5.**
*Demographic Distributions for Profiles across Subgroups in Percent.*

| Subgroup | Profile | *N* | Female | Male | American Indian | Asian | Black | Hispanic | Pacific Islander | White | Other |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | High | 6020 | 54.5 | 45.4 | 3.8 | 2.9 | 42.4 | 14.7 | 1.0 | 58.1 | 1.0 |
| | Medium | 4316 | 50.2 | 49.7 | 4.8 | 2.0 | 48.5 | 17.6 | 1.3 | 50.0 | 0.9 |
| *L_Ach* | Medium-RS | 3968 | 60.4 | 39.6 | 4.2 | 2.0 | 52.7 | 18.4 | 1.2 | 47.1 | 1.0 |
| | Low | 4278 | 47.5 | 52.3 | 5.0 | 1.8 | 53.2 | 17.6 | 1.5 | 45.1 | 1.3 |
| | Low-MW | 3706 | 48.2 | 51.7 | 5.2 | 1.6 | 58.6 | 20.3 | 1.2 | 39.7 | 1.3 |
| | HAHA | 1610 | 50.6 | 48.9 | 3.7 | 0.7 | 47.1 | 11.1 | 0.9 | 54.8 | 1.1 |
| | High | 2035 | 50.7 | 49.1 | 3.8 | 1.3 | 49.9 | 15.5 | 0.9 | 49.3 | 1.6 |
| *L_Abl* | HAMLA | 846 | 47.4 | 52.4 | 4.1 | 0.9 | 52.1 | 10.4 | 0.8 | 48.8 | 1.7 |
| | Medium | 2445 | 46.3 | 53.5 | 4.3 | 1.5 | 56.9 | 18.2 | 1.4 | 42.2 | 1.1 |
| | Low | 1714 | 41.1 | 58.7 | 4.7 | 1.5 | 56.5 | 14.6 | 1.2 | 42.1 | 1.4 |
| | High | 469 | 31.6 | 68.4 | 1.3 | 8.3 | 4.9 | 4.7 | 0.2 | 90.0 | 1.3 |
| *U_Ach* | Medium | 2383 | 36.9 | 63.0 | 1.3 | 7.1 | 7.9 | 3.6 | 0.6 | 88.7 | 1.0 |
| | Low | 2821 | 42.1 | 57.9 | 1.6 | 6.1 | 12.8 | 6.3 | 0.5 | 85.6 | 0.8 |
| | High-RS | 1397 | 58.2 | 41.7 | 1.9 | 6.4 | 13.1 | 6.3 | 0.4 | 85.1 | 1.0 |
| *U_Abl* | High-QS | 3475 | 32.7 | 67.2 | 1.9 | 9.3 | 10.6 | 7.1 | 0.7 | 83.0 | 1.2 |
| | Medium | 5225 | 44.3 | 55.6 | 2.6 | 5.8 | 19.0 | 10.8 | 0.7 | 78.6 | 1.0 |
| | Low | 2256 | 48.3 | 51.6 | 2.0 | 4.5 | 23.4 | 10.3 | 0.7 | 76.2 | 0.8 |

Note: Medium-RS = Medium Performance with Reading Strength; Low-MW = Low Performance with Math Weakness; HAHA = High Achievement High Ability (Ach>Abl); HAMLA = High Achievement Medium/Low Ability (Ach>Abl); High-QS = High Performance with Quantitative Strength; High-RS = High Performance with Reading Strength.

the HAMLA and the Low profiles. Students in the HAMLA profile could be considered "over-achievers" in Math given their potential in Quantitative ability. Strategies mitigating the risk of possible burnout may be beneficial for them to continue to excel in Math. The Quantitative and Verbal domains demonstrated the greatest variability across profiles, indicating that these areas were particularly sensitive in distinguishing differences among the latent profiles. Targeted strategies that address variability in quantitative and verbal domains could yield significant improvements.

Building on the distinctions between low-achieving and low-ability groups, a similar analysis was conducted for high-achieving and high ability groups to examine whether the derived profiles exhibit configural differences. The number of derived profile classes and profile patterns for high achieving and high ability groups differed. The profiles in *U-Ach* provided a more general categorization of performance levels (High, Medium, Low), while the *U-Abl* subgroup introduced nuanced distinctions within high-performing profiles, revealing more specific patterns of strength (High-Reading Strength, High-Quantitative Strength). All profiles within the *U-Ach* subgroup demonstrated "over-achievement" in mathematics relative to their potential in quantitative reasoning. Conversely, three profiles within the *U-Abl* subgroup were characterized by "under-achievement" in mathematics whereas the High-RS profile of this subgroup exhibited "over-achievement" in reading. This indicates that the underlying characteristics and interactions between performance metrics differ depending on whether the group is defined by achievement outcomes or inherent ability measures at the upper percentile students as well. The additional granularity in the *U-Abl* subgroup suggests more targeted interventions or instructional strategies based on domain-specific strengths.

Demographic distributions for the latent profiles across subgroups are provided in Table 5. According to the table, for both *L-Ach* and *L-Abl* subgroups, higher-performing profiles (High, Medium) show less demographic diversity than low-performing profiles, which had higher representation from underrepresented groups (Black and Hispanic students). Female representation was higher in high-performing profiles while male representation dominated in most low-performing profiles. Specifically, in the *L-Ach* subgroup, the Medium-RS profile was predominantly composed of female students, whereas the Low-MW profile was primarily comprised of male students. Both profiles, however, were significantly represented by individuals from underrepresented demographic groups, specifically Black and Hispanic students. Gender and demographic differences suggest that these factors may play a role in shaping the latent profiles in the *L-Ach* subgroup and could influence the design of targeted educational support.

For both *U-Ach* and *U-Abl* subgroups, almost all profiles were male and White-dominated. High-RS profile of *U-Abl* was an exception to this as it was dominated by females. Furthermore, higher-performing profiles were less diverse, with higher White representation and fewer underrepresented groups.

Female representation was higher in Reading-specific profiles, such as Medium-RS of *L-Ach* and High-RS of *U-Abl*, while male representation dominates in the Quantitative-specific profiles, like High-QS of *U-Abl*. Regardless of the conditions, low-performing profiles in both achievement and ability-based subgroups consistently had higher proportions of Black and Hispanic students. Gender and demographic differences indicate that these factors are likely to contribute to the formation of latent profiles and may significantly impact the development of tailored educational plans and support strategies.

The analysis also explored how the patterns of test and skill level performances compare across student profiles. In general, high-, medium-, and low-performing profiles were identified for each condition, highlighting variations among "over-achievers" (*U-Ach*, *L-Abl*) and "under-achievers" (*U-Abl*, *L-Ach*) based on mathematics achievement and quantitative reasoning. The latent profiles in the *L-Abl* subgroup showed more variations in terms of test performance than the others.

Specifically, in the low-achieving group, students exhibited notably weaker performance in mathematics relative to their quantitative reasoning skills, with this trend particularly evident in the Medium-RS and Low-MW profiles. On the other hand, students in the HAMLA profile of low ability group can be classified as "over-achievers" in mathematics given their quantitative ability performance. Within the *U-Ach* subgroup, all profiles displayed "over-achievement" in mathematics compared to their quantitative reasoning abilities. On the other hand, three profiles in the *U-Abl* subgroup showed "under-achievement" in mathematics, while the High-RS profile stood out with "over-achievement" in reading.

Figures 6 and 7 display Mathematics skill scores (percent correct scores), as well as national averages of skill scores, across the profiles of *L-Ach* and *L-Abl* subgroups, respectively. Students across the profiles of both *L-Ach* and *L-Abl* showed similar weaknesses and strengths patterns of Mathematics skills with the national sample but in varying degrees. For instance, Algebraic Patterns and Geometry were consistently strong areas whereas Measurement and Data Analysis areas showed the steepest decline across profiles in both groups. It is noteworthy that as the profiles shift from higher to lower performance levels, geometry skills increasingly dominate over algebraic pattern skills. In contrast, within the higher-performing profiles,

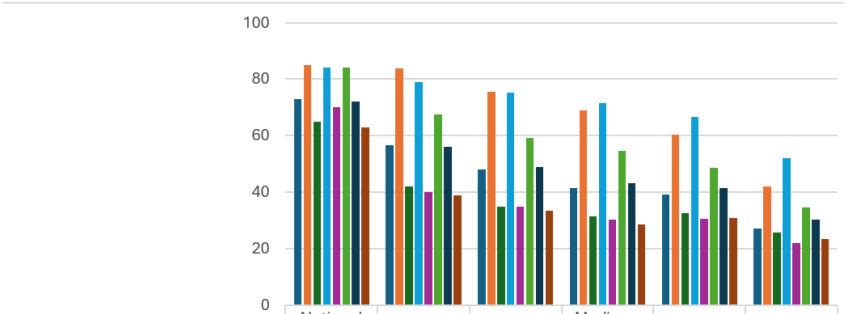
algebraic patterns skills are either comparable to or exceed those of geometry, highlighting a distinct shift in skill emphasis across performance tiers. Scores on the Extended Reasoning skill, on the other hand, were generally low, indicating this is a challenging area for all groups.

Figures 8 and 9 illustrate a comparison of skill scores across profiles of *U-Ach* and *U-Abl* relative to the national average in various mathematical domains. Consistent patterns of strengths and weaknesses were observed across profiles in both groups. Notably, all profiles within the *U-Ach* group outperformed the national averages, whereas Measurement and Extended Reasoning and to some extent the Data Analysis/Prob/Stats skill emerged as persistent challenges in the Medium and Low profiles of the *U-Abl* group. This observation highlights that high quantitative reasoning ability does not necessarily translate into high performance across all areas of mathematical achievement. Targeted efforts to address these areas of difficulty could contribute to reducing performance disparities among students. Patterns of Reading skill scores observed across profiles and conditions were more consistent; therefore, the related plots are provided in the appendix (See Figures A1-A4).

## Discussion

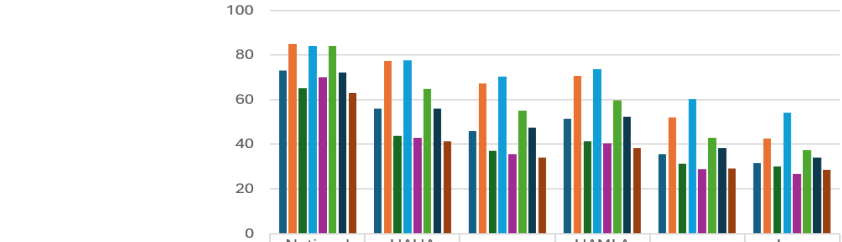
The findings highlight substantial differences in the number and patterns of latent profiles across low-achieving (*L-Ach*), low-ability (*L-Abl*), high-achieving (*U-Ach*), and high-ability (*U-Abl*) groups, emphasizing the distinct dynamics between achievement and ability, and reinforcing the notion that achievement and ability represent distinct but related constructs. Moreover, regardless of performance levels, the variations in the latent profiles between ability- and achievement-based groups support previous findings that different tests (Carman et al., 2019) and selection criteria (e.g., Lohman & Renzulli, 2007; McBee et al., 2014; Lakin, 2018) used to categorize students based on performance yield groups with distinct instructional

**Figure 6.**
*Math Skill Scores of L-Ach Subgroup with National Averages.*




| | National Avg. | High | Medium | Medium-RS | Low | Low-MW |
|---|---|---|---|---|---|---|
| Number Sense/Operations | 73 | 56,6 | 48,0 | 41,5 | 39,0 | 27,1 |
| Algebraic Patterns/Connections | 85 | 83,8 | 75,6 | 68,9 | 60,2 | 41,9 |
| Data Analysis/Prob./Stats | 65 | 41,9 | 34,8 | 31,4 | 32,7 | 25,6 |
| Geometry | 84 | 79,0 | 75,2 | 71,6 | 66,5 | 52,1 |
| Measurement | 70 | 40,1 | 34,8 | 30,3 | 30,6 | 22,0 |
| Essential Competencies | 84 | 67,5 | 59,1 | 54,6 | 48,5 | 34,6 |
| Conceptual Understanding | 72 | 55,9 | 48,9 | 43,1 | 41,5 | 30,2 |
| Extended Reasoning | 63 | 38,9 | 33,4 | 28,5 | 30,7 | 23,4 |

**Figure 7.**
*Math Skill Scores for L-Abl Subgroups with National Averages.*



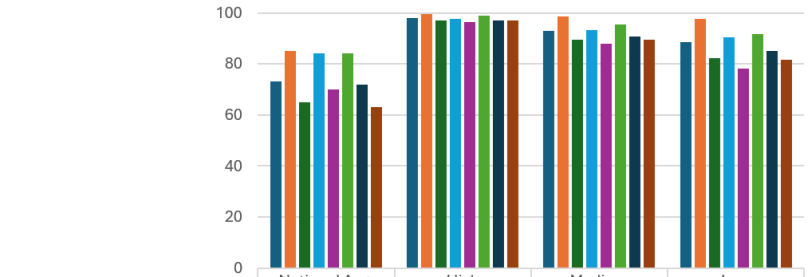

| | National Avg. | HAHA (Ach>Abl) | High | HAMLA (Ach>Abl) | Medium | Low (Ach>Abl) |
|---|---|---|---|---|---|---|
| Number Sense/Operations | 73 | 56,1 | 45,9 | 51,4 | 35,5 | 31,7 |
| Algebraic Patterns/Connections | 85 | 77,3 | 67,1 | 70,6 | 51,9 | 42,4 |
| Data Analysis/Prob./Stats | 65 | 43,9 | 36,9 | 41,5 | 31,1 | 30,1 |
| Geometry | 84 | 77,7 | 70,3 | 73,8 | 60,3 | 54,2 |
| Measurement | 70 | 42,9 | 35,5 | 40,4 | 28,9 | 26,7 |
| Essential Competencies | 84 | 64,9 | 55,1 | 59,8 | 42,8 | 37,4 |
| Conceptual Understanding | 72 | 55,9 | 47,3 | 52,2 | 38,2 | 34,1 |
| Extended Reasoning | 63 | 41,3 | 34,1 | 38,4 | 29,2 | 28,4 |

needs, especially in the gifted/talented identification. How you identify determines who you identify (Long et al., 2024).

In general, profiles in the both *L-Abl* and *U-Ach* groups had students exhibited "over-achievement" in mathematics despite lower quantitative reasoning ability is aligned with previous findings on "overachievers", who compensate for lower cognitive ability with higher perseverance, motivation, or access to enriched learning environments (Hofer & Stern, 2016; Ziernwald et al., 2022). Additionally, both the *L-Ach* and *U-Abl* groups had profiles, where mathematics performance lagged behind quantitative reasoning potential, highlighting the possible influence of external factors, instructional quality, and socioemotional barriers on student performance. Ziernwald et al. (2022) similarly reported that fluid intelligence alone does not always predict high academic performance, as motivational-affective factors and educational support structures play a crucial role in the realization of academic potential. Overall, depending on the performance level (Lower vs. Upper) of classification, achievement-based classification often overlooks cognitive potential or vice versa. This finding supports the strong recommendation of the National Association for Gifted Children (NAGC, 2010) for the use of multiple measures, especially when high stakes, test-based decisions are being made such as classroom assignment.

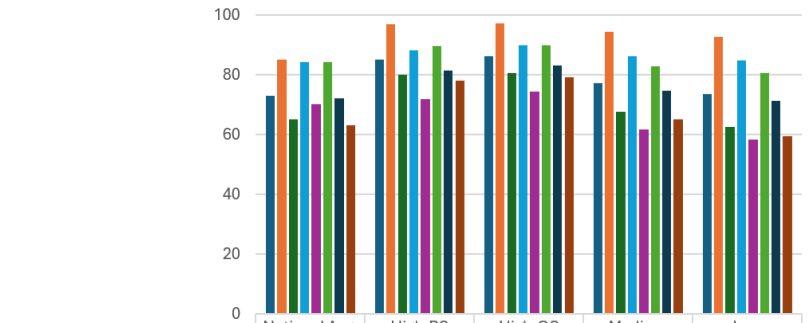
The presence of greater nuance in *U-Abl* profiles, where students displayed domain-specific strengths such as High-Reading Strength (*High-RS*) and High-Quantitative Strength (*High-QS*), as well as the diverse profiles emerged in the other groups, displayed heterogeneity in those clusters and thus the needs of differentiated instructions for the emerged profiles. This is in line with the findings that low- and high-ability students showed a larger intraindividual heterogeneity in ability indicators compared to average-ability students (Lohman et al., 2008)

**Figure 8.**
*Math Skill Scores for U-Ach Subgroup with National Averages.*




| | National Avg. | High | Medium | Low |
|---|---|---|---|---|
| Number Sense/Operations | 73 | 97,9 | 93,0 | 88,7 |
| Algebraic Patterns/Connections | 85 | 99,6 | 98,7 | 97,6 |
| Data Analysis/Prob./Stats | 65 | 97,1 | 89,6 | 82,1 |
| Geometry | 84 | 97,7 | 93,3 | 90,3 |
| Measurement | 70 | 96,5 | 87,8 | 78,1 |
| Essential Competencies | 84 | 99,0 | 95,4 | 91,6 |
| Conceptual Understanding | 72 | 96,9 | 90,8 | 85,1 |
| Extended Reasoning | 63 | 97,1 | 89,5 | 81,8 |

**Figure 9.**
*Math Skill Scores for U-Abl Subgroup with National Averages.*




| | National Avg. | High-RS | High-QS | Medium | Low |
|---|---|---|---|---|---|
| Number Sense/Operations | 73 | 85,0 | 86,1 | 77,2 | 73,5 |
| Algebraic Patterns/Connections | 85 | 96,8 | 97,0 | 94,4 | 92,4 |
| Data Analysis/Prob./Stats | 65 | 79,9 | 80,5 | 67,5 | 62,4 |
| Geometry | 84 | 88,2 | 89,6 | 86,2 | 84,8 |
| Measurement | 70 | 71,7 | 74,2 | 61,6 | 58,3 |
| Essential Competencies | 84 | 89,4 | 89,7 | 82,8 | 80,4 |
| Conceptual Understanding | 72 | 81,4 | 83,1 | 74,5 | 71,2 |
| Extended Reasoning | 63 | 77,8 | 79,0 | 65,0 | 59,5 |

Gender distribution analysis of the profiles of each group showed that female representation was higher in Reading-specific profiles, such as *Medium-RS* of *L-Ach* and *High-RS* of *U-Abl*, while male representation dominates in the Quantitative-specific profiles, like *High-QS* of *U-Abl*. This is in accordance with the long history of gender achievement gap in reading (favoring females) and math (favoring males) in the US (e.g., Robinson et al., 2011).

Demographic patterns further underscored systemic inequities, with underrepresented groups (e.g., Black and Hispanic students) predominantly occupying lower-performing profiles across all subgroups, while higher-performing profiles were less diverse and primarily composed of White students. This is consistent with the finding that the type of assessment used to categorize students had only a minor effect on equity (Hodges et al., 2018; Long et al., 2024). These findings suggest the need for interventions that are both domain-specific and equity-focused, targeting disparities in mathematics achievement and quantitative reasoning while also addressing demographic disparities to ensure more inclusive academic success.

## Conclusions

This study compared latent profiles derived from student subgroups of varying levels of mathematical skills defined by achievement and ability assessment scores. Achievement and ability cut scores for identifying students at both ends of the mathematics spectrum were applied and the resulting latent profiles within each condition were compared. The best-fitting solution across conditions ranged from 3 to 5 mutually exclusive profile classes that adequately described the variation in the ability and achievement test scores. Varying demographics and patterns of ability and achievement for each condition demonstrate the importance of recognizing students with varying learning styles and the importance of understanding distinct dynamics between achievement and ability scores while using them to identify students who may benefit from targeted instruction or placement in gifted and talented programs.

As schools continue to recover from the impact due to the disruption of the pandemic, efforts to adapt instructional strategies are crucial for ensuring students return to the pre-pandemic learning trajectory. By determining the profile characteristics, findings from this study provide valuable feedback to educators to address areas of greatest need for differentiated instruction and leveraging information regarding student academic profiles.

The LPA method used in this study enhances findings from variable-centered approaches; however, it is important to acknowledge several limitations. First, LPA does not identify "true" subgroups of individuals. Like latent variables, which are inferred from observed variables, the subgroups themselves are unobserved constructs. To address this limitation, we carefully evaluate model fit indices and examine the probabilities of each observation belonging to a given latent profile. Even though the emerged profiles across conditions allowed us to make interpretations like "over" or "under" achievement based on the ability and achievement comparison, LPA was fundamentally used as an exploratory analytical technique. This necessitates caution in drawing definitive interpretations or implications from the findings.

Despite these limitations, this study represents an important exploratory step in identifying potential unique profiles of second graders' achievement and ability performances. The current study is based on one large educational system; therefore, the generalization of the results might be limited. Future research should explore whether these profiles replicate across different populations and settings to validate and extend the current findings. Students interpret their experiences through a combination of cognitive, social, and emotional processes, all of which impact learning (Darling-Hammond & Cook-Harvey, 2018). Given that, one should investigate the connections among them in terms of identifying potential unique profiles. Furthermore, a multiple-group latent profile analysis (Morin, et al., 2016) should be conducted to make direct comparisons within conditions used in this study to investigate the invariance of emerged profiles.

## References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Akogul, S., & Erisoglu, M. (2017). An Approach for Determining the Number of Clusters in a Model-Based Cluster Analysis. *Entropy*, *19*(9), 452. MDPI AG. http://dx.doi.org/10.3390/e19090452.

Banfield, J.D., & Raftery, A.E (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics,* 49, 803–821.

Biernacki, C., & Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics, 29*, 451–457.

Carman, C. A., Walther, C. A. P., & Bartsch, R. (2019, April 6). Effects of test selection on the gifted identification process. Paper presented at the 2019 annual meeting of the American Educational Research Association, Toronto, Canada. Retrieved [02/01/2025], from AERA Online Paper Repository. https://doi.org/10.3102/1437767

Carriedo, N., Rodríguez-Villagra, O. A., Pérez, L., & Iglesias-Sarmiento, V. (2024). Executive functioning profiles and mathematical and reading achievement in Grades 2, 6, and 10. *Journal of School Psychology, 106*, 101353. https://doi.org/10.1016/j.jsp.2024.101353

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1–22.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. Amsterdam: North-Holland.

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters, 42*(4), 333–343. https://doi.org/10.1016/S0167-7152(98)00200-4.

Curriculum Associates, LLC. (2020). *Understanding Student Needs: Early Results from Fall Assessments*. Curriculum Associates Research Report No. 2020-40. North Billerica, MA: Author.

Darling-Hammond, L., & Cook-Harvey, C. M. (2018). *Educating the whole child: Improving school climate to support student success.* Retrieved from Learning Policy Institute https://learningpolicyinstitute.org/media/547/download?inline&file=Educating_Whole_Child_REPORT.pdf

Dunbar, S. B., & Welch, C. J. (2015). *Iowa Assessments, Form E*. Riverside Insights: Itasca, IL.

Green, C. T., Bunge, S. A., Chiongbian, V. B., Barrow, M., & Ferrer, E. (2017). Fluid reasoning predicts future mathematical performance among children and adolescents. *Journal of Experimental Child Psychology, 157*, 125–143. https://doi.org/10.1016/j.jecp.2016.12.005

Hart, S. A., Logan, J. A. R., Thompson, L., Kovas, Y., McLoughlin, G., & Petrill, S. A. (2016). A latent profile analysis of math achievement, numerosity, and math anxiety in twins. *Journal of Educational Psychology, 108*(2), 181–193. https://doi.org/10.1037/edu0000045

Hodges, J., Tay, J., Maeda, Y., & Gentry, M. (2018). A meta-analysis of gifted and talented identification practices. *Gifted Child Quarterly, 62*(2), 147–174. https://doi.org/10.1177/0016986217752107

Hofer, S. I., & Stern, E. (2016). Underachievement in physics: When intelligent girls fail. Learning and Individual Differences, 51, 119–131. https://doi.org/10.1016/j.lindif.2016.08.006

Hwang, J., Hand, B., & French, B. F. (2023). Critical thinking skills and science achievement: A latent profile analysis. *Thinking Skills and Creativity, 49*, 101349. https://doi.org/10.1016/j.tsc.2023.101349

Jesson, R. (2018). Stanines. In The SAGE encyclopedia of educational research, measurement, and evaluation (Vol. 4, pp. 1610-1611). SAGE Publications, Inc., https://doi.org/10.4135/9781506326139

Kuhfeld, M., Tarasawa, B., Johnson, A., Ruzek, E., & Lewis, K. (2020). *Learning during COVID-19: initial findings on students' reading and math achievement and growth*. NWEA.

Lakin, J. M. (2018). Making the cut in gifted selection: Score combination rules and their impact on program diversity. Gifted Child Quarterly, 62(2), 210-219. https://doi.org/10.1177%2F0016986217752099

Laursen, B., & Hoff, E. (2006). Person-centered and variable-centered approaches to longitudinal data. *Merrill-Palmer Quarterly (1982-)*, 377–389.

Lubke, G., & Neale, M. C. (2006). Distinguishing Between Latent Classes and Continuous Factors: Resolution by Maximum Likelihood? *Multivariate Behavioral Research*, *41*(4), 499–532. https://doi.org/10.1207/s15327906mbr4104_4.

Lohman, D. F. (2013) *CogAT Form 7 Score Interpretation Guide: Part 3 Adapting Instruction to Student's Needs and Abilities*. Riverside Insights.

Lohman, D. F., Gambrell, J., & Lakin, J. (2008). The commonality of extreme discrepancies in the ability profiles of academically gifted students. *Psychology Science Quarterly, 50*(2), 269–282.

Lohman, D. F., & Lakin, J. M. (2017). *Cognitive Abilities Test (CogAT), Form 7*. Riverside Insights.

Lohman, D. F., & Renzulli, J. (2007). A simple procedure for combining ability test scores, achievement test scores, and teacher ratings to identify academically talented children. Retrieved from http://faculty.education.uiowa.edu/docs/dlohman/Lohman_ Renzulli_ID_system.pdf

Long, D. A., Peters, S., McCoach, D. B., & Gambino, A. J. (2024). How you identify determines who you identify: The implications of the choice of talent measures, norms, cut-offs, and combination rules on the academic profile and diversity of students identified as gifted. OSF Preprints. https://doi.org/10.31219/osf.io/rf9wn

Litkowski, E. C., Finders, J. K., Borriello, G. A., Purpura, D. J., & Schmitt, S. A. (2020). Patterns of heterogeneity in kindergarten children's executive function: Profile associations with third grade achievement. *Learning and Individual Differences, 80*, 101846. https://doi.org/10.1016/j.lindif.2020.101846

Mahatmya, D., Assouline, S., Foley-Nicpon, M., Ali, S. R., McGinnis, D., & Teriba, A. (2023). Patterns of high ability and underrepresented students' subject-specific psychosocial strengths: A latent profile analysis. *High Ability Studies*, *34*(2), 229–248. https://doi.org/10.1080/13598139.2023.2176293

Mammadov, S., Ward, T. J., Cross, J. R., & Cross, T. L. (2016). Use of latent profile analysis in studies of gifted students. *Roeper Review, 38*(3), 175–184. https://doi.org/10.1080/02783193.2016.1183739

Marsh, H. W., Lüdtke, O., Trautwein, U., & Morin, A. J. (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person-and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling, 16*(2), 191–225.

McBee, M. T., Peters, S. J., & Waterman, C. (2014). Combining scores in multiple-criteria assessment systems: The impact of combination rule. *Gifted Child Quarterly, 58*(1), 68-89. https://doi.org/10.1177%2F0016986213513794

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. John Wiley & Sons. https://doi.org/10.1002/0471721182

Moon, S. M., & Reis, S. M. (2004) Acceleration and twice exceptional students. In N. Colangelo, S. G. Assouline, & M. U. M. Gross (Eds.), *A nation deceived: How schools hold back America's brightest students* (Vol. 2, pp. 109-119). Iowa City, IA: The Connie Belin & Jaqueline N. Blank International Center for Gifted Education and Talent Development.

Morin, A., Meyer, J., Creusier, J., & Biétry, F. (2016). Multiple-group analysis of similarity in latent profile solutions. *Organizational Research Methods, 19*(2), 231–254. https://doi.org/10.1177/1094428115621148

Moses, R. P., & Cobb, C. E. Jr., (2001). *Radical equations: Civil rights from Mississippi to the Algebra Project*. Boston: Beacon

National Association for Gifted Children. (2010). National standards in gifted and talented education: Pre-K to Grade 12 gifted education programming standards. Retrieved from http://www.nagc.org/sites/default/files/standards/K-12%20programming%20standards.pdf

National Center for Education Statistics. (2019). *NAEP, Reading Report Card for the Nation and the States*. U.S. Department of Education.

Nickerson, R. (2011). Developing intelligence through instruction. In R. J. Sternberg, & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 107–129). New York: Cambridge University Press.

R Core Team. (2023). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. https://www.R-project.org

Renaissance Learning (2021). *How kids are performing: Tracking the school-year impact of COVID-19 on reading and mathematics achievement*. Special Report Series, Spring 2021 edition.

Riverside Insights. (2012). *Forms E and F Score Interpretation Guide*. Itasca, IL: The University of Iowa (Iowa Testing Program).

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. American Educational Research Journal, 48(2), 268–302. https://doi.org/10.3102/0002831210372249

Rosenberg, J. M., van Lissa, C. J., Beymer, P. N., Anderson, D. J., Schell, M. J. & Schmidt, J. A. (2019). tidyLPA: Easily carry out Latent Profile Analysis (LPA) using open-source or commercial software [R package]. https://data-edu.github.io/tidyLPA/

Schneider, W. (2013). Principles of assessment of aptitude and achievement. In D. H. Saklofske, V. L. Schwean, & C. R. Reynolds (Eds.), The Oxford handbook of child psychological assessment (pp. 286–330). New York: Oxford

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6(2)*, 461–464.

Soares, D. L., Lemos, G. C., Primi, R., & Almeida, L. S. (2015). The relationship between intelligence and academic achievement throughout middle school: The role of students' prior academic performance. *Learning and Individual Differences, 41,* 73–78. https://doi.org/10.1016/j.lindif.2015.02.005

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior, 120*, 103445. https://doi.org/10.1016/j.jvb.2020.103445

Younger, J. W., Schaerlaeken, S., Anguera, J. A., & Gazzaley, A. (2024). The whole is greater than the sum of its parts: Using cognitive profiles to predict academic achievement. *Trends in Neuroscience and Education, 36,* 100237. https://doi.org/10.1016/j.tine.2024.100237

Ziernwald, L., Schiepe-Tiska, A., & Reiss, K. M. (2022). Identification and characterization of high-achieving student subgroups using two methodological approaches: The role of different achievement indicators and motivational-affective characteristics. *Learning and Individual Differences, 100*, 102212. https://doi.org/10.1016/j.lindif.2022.102212

## Appendix A

**Table A1.**

*Skill Definition Table for the Iowa Assessments.*

| Subject | Skill Domain Description |
|---|---|
| Reading | Conceptual Understanding<br>Essential Competencies<br>Extended Reasoning<br>Literary<br>Explicit Meaning<br>Implicit Meaning<br>Informational<br>Key Ideas |
| Mathematics | Algebraic Patterns & Connections<br>Conceptual Understanding<br>Essential Competencies<br>Extended Reasoning<br>Geometry<br>Measurement<br>Number Sense & Operations<br>Data Analysis, Probability, & Statistics |

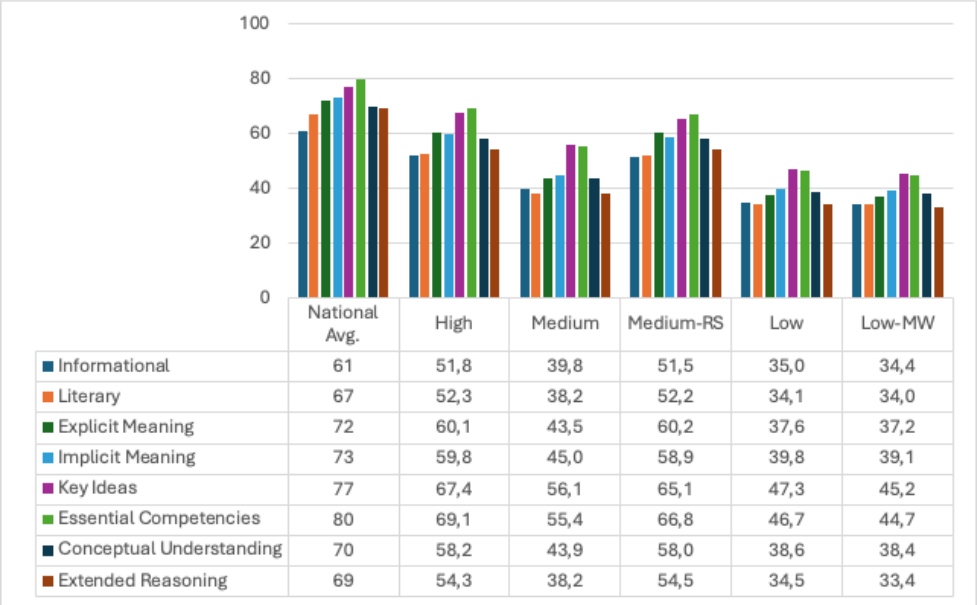
**Table A2.**

*Alignment by Subject of Tests and Standards for the Iowa Assessments.*

| Subject | Alignment with Standards |
|---|---|
| Reading | National Council of Teachers of English (NCTE) and International Reading Association (IRA) Standards for the English Language Arts |
| Mathematics | National Council of Teachers of Mathematics (NCTM) Assessment Standards for School Mathematics; Curriculum and Evaluation Standards for Mathematics |

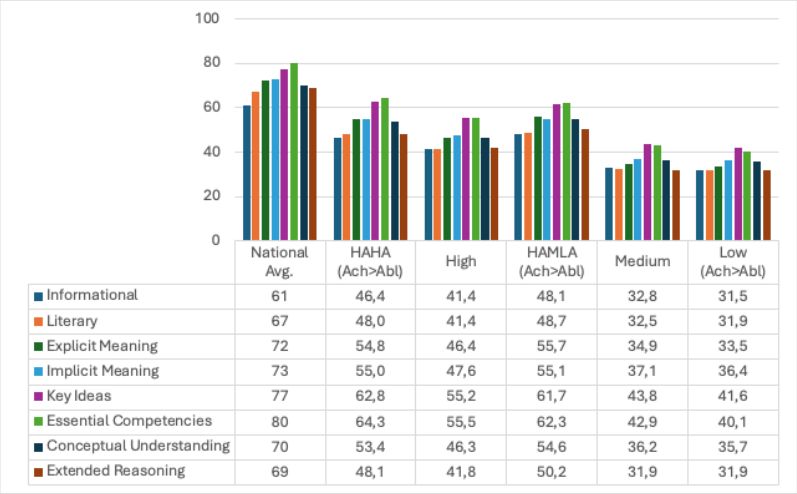
**Figure A1.**

*Reading Skill Scores of L-Ach Subgroup with National Averages.*



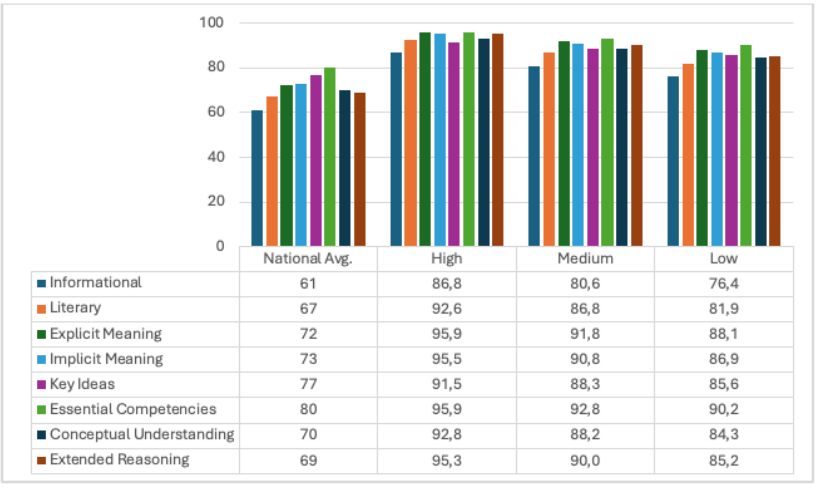

| | National Avg. | High | Medium | Medium-RS | Low | Low-MW |
|---|---|---|---|---|---|---|
| Informational | 61 | 51,8 | 39,8 | 51,5 | 35,0 | 34,4 |
| Literary | 67 | 52,3 | 38,2 | 52,2 | 34,1 | 34,0 |
| Explicit Meaning | 72 | 60,1 | 43,5 | 60,2 | 37,6 | 37,2 |
| Implicit Meaning | 73 | 59,8 | 45,0 | 58,9 | 39,8 | 39,1 |
| Key Ideas | 77 | 67,4 | 56,1 | 65,1 | 47,3 | 45,2 |
| Essential Competencies | 80 | 69,1 | 55,4 | 66,8 | 46,7 | 44,7 |
| Conceptual Understanding | 70 | 58,2 | 43,9 | 58,0 | 38,6 | 38,4 |
| Extended Reasoning | 69 | 54,3 | 38,2 | 54,5 | 34,5 | 33,4 |

**Figure A2.**
*Reading Skill Scores for L-Abl Subgroups with National Averages.*



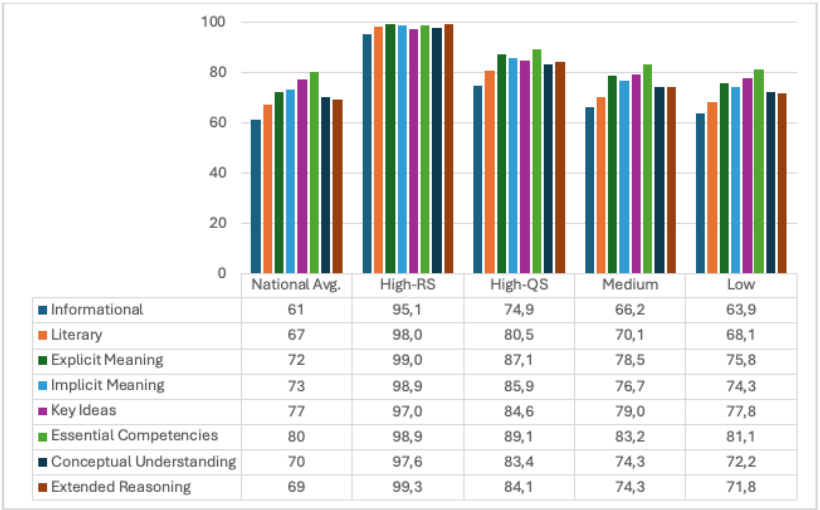

| | National Avg. | HAHA (Ach>Abl) | High | HAMLA (Ach>Abl) | Medium | Low (Ach>Abl) |
|---|---|---|---|---|---|---|
| Informational | 61 | 46,4 | 41,4 | 48,1 | 32,8 | 31,5 |
| Literary | 67 | 48,0 | 41,4 | 48,7 | 32,5 | 31,9 |
| Explicit Meaning | 72 | 54,8 | 46,4 | 55,7 | 34,9 | 33,5 |
| Implicit Meaning | 73 | 55,0 | 47,6 | 55,1 | 37,1 | 36,4 |
| Key Ideas | 77 | 62,8 | 55,2 | 61,7 | 43,8 | 41,6 |
| Essential Competencies | 80 | 64,3 | 55,5 | 62,3 | 42,9 | 40,1 |
| Conceptual Understanding | 70 | 53,4 | 46,3 | 54,6 | 36,2 | 35,7 |
| Extended Reasoning | 69 | 48,1 | 41,8 | 50,2 | 31,9 | 31,9 |

**Figure A3.**
*Reading Skill Scores for U-Ach Subgroup with National Averages.*




| | National Avg. | High | Medium | Low |
|---|---|---|---|---|
| Informational | 61 | 86,8 | 80,6 | 76,4 |
| Literary | 67 | 92,6 | 86,8 | 81,9 |
| Explicit Meaning | 72 | 95,9 | 91,8 | 88,1 |
| Implicit Meaning | 73 | 95,5 | 90,8 | 86,9 |
| Key Ideas | 77 | 91,5 | 88,3 | 85,6 |
| Essential Competencies | 80 | 95,9 | 92,8 | 90,2 |
| Conceptual Understanding | 70 | 92,8 | 88,2 | 84,3 |
| Extended Reasoning | 69 | 95,3 | 90,0 | 85,2 |

**Figure A4.**
*Reading Skill Scores for U-Abl Subgroup with National Averages.*




| | National Avg. | High-RS | High-QS | Medium | Low |
|---|---|---|---|---|---|
| Informational | 61 | 95,1 | 74,9 | 66,2 | 63,9 |
| Literary | 67 | 98,0 | 80,5 | 70,1 | 68,1 |
| Explicit Meaning | 72 | 99,0 | 87,1 | 78,5 | 75,8 |
| Implicit Meaning | 73 | 98,9 | 85,9 | 76,7 | 74,3 |
| Key Ideas | 77 | 97,0 | 84,6 | 79,0 | 77,8 |
| Essential Competencies | 80 | 98,9 | 89,1 | 83,2 | 81,1 |
| Conceptual Understanding | 70 | 97,6 | 83,4 | 74,3 | 72,2 |
| Extended Reasoning | 69 | 99,3 | 84,1 | 74,3 | 71,8 |

iejee

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

# Exploring Test-Taking Disengagement in the Context of PISA 2022: Evidence from Process Data

Başak Erdem Kara[a,*]

Received : 13 December 2024
Revised : 10 January 2025
Accepted : 2 March 2025
DOI : 10.26822/iejee.2025.380

[a,*] **Corresponding Author:** Başak Erdem Kara, Anadolu University, Education Faculty, Department of Educational Sciences, Eskisehir, Türkiye.
E-mail: basakerdem@anadolu.edu.tr
ORCID: https://orcid.org/0000-0003-3066-2892

## Abstract

Achievement tests are commonly used in education to evaluate students' academic performance and proficiency in specific subject areas. However, there is a major problem that threatens the validity of achievement test scores which is test-taking disengagement. Respondents provide answers that are inconsistent with their true ability level and can introduce construct irrelevant variance that threatens the validity of scores. This study examines test-taking disengagement in the context of PISA 2022 using process data to identify patterns of behavior that influence student performance. Three key indicators; response time, number of actions and self-reported effort, were used to examine engagement levels. Employing Latent Profile Analysis (LPA), distinct profiles of test-takers were identified, ranging from highly engaged to disengaged groups. Results indicate that disengagement, characterized by low self-reported effort, minimal interactions, and rapid responses, is associated with lower test performance, threatening the validity of scores. These findings highlight the significance of accounting for disengagement when interpreting the results of large-scale assessments. The implications were discussed in relation to the existing literature and recommendations for future research were provided to address identified gaps and extend the study's contributions.

**Keywords:**

Test-Taking Disengagement, Response Time, Number Of Action, Self-Reported Effort

## Introduction

Achievement tests are a widely used tool in education to assess student performance, with the primary intention of measuring what a student knows and can do when they are fully engaged and demonstrating their maximum performance while responding to items (Cronbach, 1960; Messick 1989). Ideally, students are assumed to exert maximum effort on test items, ensuring that test scores accurately reflect the construct being measured. In practice, however, this ideal scenario is not always achieved, as some students may not put forth the effort necessary to thoroughly process an item and provide responses that are consistent with their true ability (Wise, 2017; Wise & Kingsburry, 2016, 2022).

It is recognized that a valid achievement test score requires an engaged test-taker demonstrating what they know and can do (Cronbach, 1960; Messick, 1984). However, test-takers may feel unmotivated to exert effort, particularly in low-stakes tests where they often believe their performance has no personal consequences. Consequently, when test-takers respond with inadequate effort, their test scores are likely to reflect a lower level of ability than they actually possess. This behavior, known as test-taking disengagement, introduces non-negligible, construct-irrelevant variance that poses a potential threat to score validity (Eklöf, 2010; Goldhammer et al., 2016; Kong et al., 2007; Wise, 2017). In general, test-taking disengagement is defined as providing responses that are inconsistent with those expected from engaged test-takers. It includes situations in which the individual provides a response without reference to his or her knowledge, skills, or abilities (Soland et al., 2019).

*Test-Taking Disengagement and PISA*

Programme for International Student Assessment (PISA) is one of the International Large Scale Assesments (ILSA) regularly administering tests and questionnaires. Its purpose is to evaluate the readiness of 15-year-old students to tackle the challenges of today's information-driven society and to draw conclusions about the effectiveness of a country's education system. The program focuses on students' ability to use their knowledge and skills to meet with real-life challenges, rather than on their mastery of a particular area of the school curriculum (OECD, 2024). In the PISA, students take a test designed to measure their skills, typically in mathematics, reading, and science. Participation is voluntary and anonymous, with minimal to no direct consequences for the students. As a result, the test is considered a low-stakes assessment at the individual respondent level (Baumert & Demmrich, 2001; Finn, 2015; Pools & Monseur, 2021).

As in other assessment situations, PISA also assumes that the scores obtained by test-takers reflect only differences in the characteristics measured, but test-takers may not give their best effort that would be desired (Buchholz et al., 2022). Thus, the validity of the inferences based on the PISA assessment needs to be controlled and demonstrated. As we discussed before, in low-stakes testing contexts, such as PISA, there are often no personal consequences for test-takers, i.e. any form of incentive, influence on academic record or feedback. Research has consistently shown that low-stakes assessments tend to produce lower levels of engagement. Disengagement is the main construct-irrelevant factor that jeopardizes the validity of low-stakes test scores, and test administrators are aware of and concerned about its potential impact (Finn, 2015; Wise, 2020; Wolf and Smith, 1995). Because PISA is also a low-stakes assessment, it is also open to the validity threat posed by disengagement.

*Indicators of Test-Taking Disengagement*

There are several measures to examine students (dis) engagement that are typically categorized as self-reported effort (SRE) data and test-takers response behavior. Response behaviors include behavioral analysis demonstrated by students while completing an assessment. In the context of ILSAs, test-based behavioral measures can be derived from either response patterns or process data collected during computer-based assessments (CBAs) (Buchholz et al., 2022). In the context of this study, log data measures and the SRE are the main focus and are discussed in detail below.

Process Data (Log Data). The use of CBAs has introduced alternative approaches leveraging log data. These assessments enable the collection of data that capture not only the answers provided by test-takers but also their observable behaviors during the test. This type of data, known as process or log data, includes metrics such as the time spent on each question, the frequency and nature of interactions, and the intervals between actions. Such data offer researcher valuable insights into both the test-takers' final responses and the cognitive processes they employed to reach those answers (Ramalingam, 2017). Recently, log file data have been utilized to identify instances of disengagement during test-taking (Gobert et al., 2015). The most widely used approach relies on the amount of time individuals spend responding to an item. These methods are based on the assumption that participants exhibiting low effort complete tasks more quickly and spend less time on them compared to those who are more motivated (Wise & Kong, 2005). Response time data is regarded as a less biased approach because it reflects actual behavior rather than self-reported evaluations and does not require any extra effort from the respondents. This approach allows for more accurate and continuous tracking of changes in engagement because response data is collected for each individual item rather than at specific points in time (Wise & Kong, 2005). In addition to response time, number of actions data from the log file could be used as complementary measure to examine disengagement. Number of actions reflects examinees' interactions with a specific item, serving as an indicator of their behavioral engagement with the task. Sahin and Colvin (2020) stated that a lower number of clicks is an indicator of lower levels of motivation and thus higher levels of disengagement.

Self-Reported Effort. One of the most widely used methods to assess engagement is to ask test-takers to directly self-report the amount of the effort they put into taking the test. For example, PISA employs an "effort thermometer" (Butler & Adams, 2007), in which

test-takers rate their engagement on a scale from 1 (lowest) to 10 (highest). Despite their ease of use, self-report measures have notable limitations. First, the accuracy of the data may be questionable because self-report measures are susceptible to response bias. Second, the interpretation of self-report scores can be challenging, as these scores may not provide clear insight into the specific nature or extent of disengagement (Wise, 2020).

*Test-Taking Disengagement and Test Performance*

As discussed before, disengaged responding introduces a construct-irrelevant variance into the measurement process and its presence threatens the interpretation of test scores which can lead to some poor decisions (Wise & Kingsburry, 2022). Previous research has consistently highlighted a relationship between test-taking effort and achievement. In general, higher levels of engagement are associated with higher levels of test performance (Kuhfeld & Soland, 2020). Motivated students tend to perform better on tests than unmotivated ones (Wise & DeMars, 2005; Wise & Kong, 2005; Finn, 2015).

In contrast, according to Gignac et al. (2019) it is not necessary to exert maximum effort or to have a very high level of test-taking motivation to obtain valid test scores but rather reaching a sufficient level of effort. While effort generally improves performance, there are exceptions such as cases where students in low-effort clusters achieved high scores, i.e., test-taking effort had a weak negative correlation with test performance (Lundgren & Eklöf, 2020). In the context of low-stakes assessments, both motivational and cognitive factors are found to explain test performance, making the interpretation of results less straightforward. Eklöf et al. (2014) show that controlling for effort changes the ranking of countries in the TIMSS results. Zamarro et al. (2019) found that effort accounted for 32-38% of the variation in PISA 2009 scores. Similarly, Akyol et al. (2021) estimated that a country could improve its PISA ranking by up to 15 places if all students took the test seriously. These findings underscore that achievement test results are shaped by both student ability and motivation.

*Present Study*

Test-taking disengagement and its relationship to test performance and psychometric properties has become an important concern and significant area of interest for researchers and practitioners due to the validity challenges it poses (Wise & DeMars, 2005; Wise, 2016). Previous studies have proposed various process data-based approaches to detect unmotivated responses; however, these methods frequently produce differing outcomes when applied to the same sample (Goldhammer et al., 2016). While test-taking effort is generally positively correlated with performance, this relationship is less clear in some studies (Gignac et al., 2019; Lundgren & Eklof, 2020). Therefore, the present research aimed to examine students' test-taking effort using various indicators, specifically self-reported effort and log data, including response time and the number of actions, within the context of the PISA 2022 dataset in the Turkish sample. The Turkish sample was selected because Turkey was one of the countries that included a measure of self-reported effort and process data records in the PISA 2022 assessment, and it also ranked among the countries with the highest test effort in the PISA 2018 cycle. Turkish students had high levels of engagement based on behavioral indicators (low non-response and rapid guessing rates) and high level of self-reported effort (Buchholz, 2022). This makes Turkey a particularly relevant context for the study. In this study, Latent Profile Analysis (LPA) is used to identify the different groups that define students' test effort. This analysis will not only provide new insights into understanding student effort levels, but will also provide a deeper understanding for accurately assessing test performance. Answers were sought to the following research questions:

> *RQ1. What percentage of the sample show disengagement?*
>
> *RQ2. How does effort, as reflected in process data (response time and number of actions), self-reported effort, and test performance, relate to one another?*
>
> *RQ3. What profiles can students be classified into based on response time, number of actions, and self-reported effort data?*

In addition, some factors such as item type, demographic characteristics of the sample, item position etc. may influence the test-taking profiles and gender was taken into consideration to examine the results of LPA in depth.

Self-Reported Effort. On the last page of the PISA assessment booklet or screen, there is a section called the PISA Effort Thermometer and students are asked to imagine a situation that they consider important and for which they would do their best and exert as much effort as possible. Students are asked to rate their self-reported effort (SRE) based on these statements using a scale of 1 to 10, with 10 being maximum effort. They are presented with the following question and asked to rate their effort (OECD, 2016).

"How much effort did you put in doing this test [PISA]?"

Here, a score of 10 indicates that students believe they put as much effort into the PISA test as they would in a real-life scenario of great importance to them (OECD, 2016).

Mathematics Performance. As mentioned above, mathematics is the main domain of the PISA 2022 assessment, so we focused on mathematical items and performance scores. The computer-based PISA 2022 assessment spanned two hours, divided into two one-hour sessions with a 5-minute break in between (OECD, 2024). Students were tasked with completing two 30-minute clusters of items in each session, amounting to four clusters in total. While two clusters were dedicated to the major domain, the remaining clusters assessed one or two of the minor domains. The PISA 2022 item pool included 99 items and a total of 234 mathematics questions (OECD, 2024).

**Data Analysis**

To obtain the response time (RT) and number of actions (NA) scores, we calculated the average RT and NA values for each individual. Missing values were excluded by listwise deletion and this cleaning process resulted in a sample of 6560 out of 7250 students. In addition to the raw scores of RT and NA scores, we also calculated an effort index to examine the frequency of disengaged responders on the sample. The response time effort (RTE) index was introduced by Wise and Kong (2005) and calculated as follows;

$$SB_{ij} = \begin{cases} 1, & if\ RT_{ij} \geq T_i \\ 0, & if\ RT_{ij} < T_i \end{cases} \qquad RTE = \frac{\sum SB_{ij}}{k}$$

In this formula, SBij refers to the solution behavior for the item i and person j and is calculated based on a threshold value (Ti). k refers to the number of items. In this point, RTE indicates the proportion of items in which solution behavior is shown. A higher value is assumed to be an indicator of greater test-taking effort and engagement during the test.

In our study, we examined two distinct thresholds: a 5-second threshold (Wise & Kong, 2005) and the normative threshold (NT10; Wise & Ma, 2012). The 5-sec threshold serves as a benchmark for the minimum time needed to meaningfully engage with an item. A response time below 5 seconds is interpreted as a sign of low effort or disengagement by the respondent. This threshold is useful for differentiating rapid guessing, where responses are made too quickly to demonstrate genuine effort, from intentional and effortful engagement (Wise & Kong, 2005). On the other hand, the NT10 threshold is defined as 10% of the average time test-takers spend on an item, at a maximum of ten seconds. We couldn't find an RTE-like formula used for NA in the literature. We adapted the RTE formula to NA based on the normative 10 method. Thus, we set our threshold by taking the 10% of the average NA that test takers had on an item, with the goal of following a similar logic to response time and ensuring consistency in the application of effort measures. However, we acknowledge that this is only an attempt to adapt the RTE formula. The threshold obtained may not be universally applicable, and further research is needed to refine these criteria. Readers should be aware of this and use and interpret the results with caution.

To examine the consistency of different measures and their relationship with achievement, Pearson correlations were examined. In addition, the presence of different subgroups of disengaged responders were investigated with LPA using the following indices: response time, number of actions, self-reported effort. Latent profile analysis (LPA) is a statistical technique used to uncover and characterize hidden groups of individuals (referred to as profiles in LPA) who exhibit similar patterns across one or more indicator variables. These groups, often referred to as unobserved latent mixture components, can be conceptualized as distinct classes or profiles of individuals. LPA falls under the broader category of Mixture Models (Ferguson et al., 2020, Hofverberg et al., 2022). Because LPA, unlike many traditional statistical methods, emphasizes the grouping of individuals rather than variables, it is often referred to as a person-centered approach to statistical analysis, as opposed to a variable-centered approach. Prior to conducting the analysis, multivariate normality was assessed using the Mardia test via the psych package in R (Revelle, 2022) in order to account for potential violations. Due to significant departures from normality, with both skewness and kurtosis showing p-values less than 0.01, the MLR estimator was chosen for its robustness to normality violations and its ability to produce more stable results (Li, 2015; Vermunt & Magidson, 2002). When using the MLR estimator, the inclusion of various fit indices contributes to a clearer interpretation and more robust model evaluation. While aBIC is particularly relevant due to its sample size adjustment, it is also important to consider other indices such as BIC, AIC and entropy when evaluating model fit and classification accuracy. Lower aBIC, BIC and AIC values indicate a better fitting model, while entropy values closer to 1 indicate a more accurate classification. In addition, likelihood ratio tests (e.g., LMR-LRT, BLRT) are useful for comparing models with different numbers of latent profiles to assess whether additional profiles significantly improve model fit (Morgan, 2015; Nylund et al., 2007; Spurk et al., 2020). Briefly, the number of groups was determined based on AIC, BIC, aBIC, entropy value, Lo-Mendell-Rubin likelihood ratio test (LMR), interpretability of the resulting groups, and the parsimony principle. Both the descriptive analysis and the LPA (using the MplusAutomation package (Hallquist & Wiley, 2018) with Mplus7 (Muthén & Muthén, 2014)) were performed in R statistical software (v2024.09.1+394; R Core Team, 2024).

**Results**

In this section, we first present descriptive statistics and correlations between different measures. Next,

we interpret the results of the latent profile analysis, including how we classified students into profiles, how we determined the optimal model, and how we described the resulting profiles. Finally, we examine the relationship between the profiles and students' mathematics achievement and effort.

What percentage of the sample show disengagement?

Table 1 presents the distribution of students' engagement across three metrics: RTE_5sec, RTE_10p, and NA_10p. Engagement is categorized as Fully Engaged (=1), Highly Engaged (>.90), Moderately Engaged (.90 - .80), and Low Engaged (<.80).

**Table 1.**
*Number of engaged and disengaged students under three different threshold system*

| | RTE_5sec | RTE_10p | NA_10p |
|---|---|---|---|
| Fully engaged (1.00) | 5735 (82.37%) | 4526 (65.00%) | 382 (5.49%) |
| Highly engaged (>.90) | 977 (14.03%) | 1782 (25.59%) | 1041 (14.95%) |
| Moderately engaged (.80 - .90) | 173 (2.48%) | 406 (5.83%) | 1919 (27.56%) |
| Low engaged (<.80) | 78 (1.12%) | 249 (3.58%) | 3621 (52.00%) |

The data in Table 1 shows that under 5-sec threshold, the number of low engaged respondents was 78 (1.12%) and the number of medium engaged respondents was 173 (2.48%). The RTE_10 percent method provided more conservative results than the common threshold method. The number of fully engaged students were fewer on this normative method. On the other hand, the number of actions methods classified most of the examinees (52.00%) as low engaged test takers. The NA_10p metric, likely reflecting a call for further investigation and try with another threshold method due to its much lower engagement distribution.

*How does effort, as reflected in process data (response time and number of actions), self-reported effort, and test performance, relate to one another?*

Descriptive statistics and Pearson correlations for each pair of measures, that have the potential to serve as indicators of disengaged responding: response time (RT), number of actions (NA), self-reported effort (SRE) and mathematics achievement (Ach), are provided in Table 2.

**Table 2.**
*Correlations and descriptive statistics between variables*

| | RT | NA | SRE | Ach | Mean | *SD* |
|---|---|---|---|---|---|---|
| Response Time (RT) | 1.00 | | | | 93.66 | 23.17 |
| Number of Actions (NA) | .37 | 1.00 | | | 20.4 | 11.47 |
| Self-Reported Effort (SRE) | .09 | .03 | 1.00 | | 8.14 | 2.12 |
| Math Achievement (Ach) | .40 | .43 | .02 | 1.00 | 452.24 | 89.29 |

The mean response time for the Turkey sample is 93.66 seconds (*SD* = 23.17) and the mean number of actions is 20.4 (*SD* = 11.47) for an item. The self-reported effort (SRE) item has an average of 8.14 out of 10 which is indicating a high level of self-effort. Lastly, the average mathematics achievement mean score is 452.24.

Notable relationships are observed between RT, NA, SRE, and mathematics achievement. To illustrate, the strongest correlation with achievement is observed for the NA ($r$ = .43). The correlation between RT and achievement is relatively low ($r$ = .40). Notably, SRE has the lowest correlation with performance, with correlation coefficients of .02. Similarly, the correlations between the SRE and RT ($r$ = .09) and number of actions ($r$ = .03) are weak, suggesting that these items may have a limited relationship with process data based methods for identifying disengaged responses. Conversely, the positive correlation between response time (RT) and the number of actions (NA) ($r$ = .37) suggests that longer RT are associated with a higher NA, which may indicate a higher level of engagement in the test taking process. These findings highlight the importance of considering RT, NA, SRE, and performance-related variables in understanding disengaged responding.

*What profiles can students be classified into based on response time, number of actions and self-reported effort data?*

In the context of this study, LPA was used to classify students into subgroups based on different measures of disengagement. As stated in the methods section, the Mardia test results revealed significant deviations from multivariate normality, with both skewness and kurtosis showing p-values less than .01. Consequently, the MLR estimator was preferred for LPA and the results of the analysis are presented in Table 3.

Table 3 shows the fit indices of the LPA models for the different profile solutions. When deciding on the optimal solution, the lower AIC, BIC and aBIC values indicate a better fit and higher entropy values indicate a higher classification confidence. The p-value of the LMR test is also taken into account. Considering all these indicators, the three-profile model was considered as the optimal solution. The model fit statistics presented in Table 3 indicate that the three-profile solution provides the optimal balance between statistical fit and interpretability. The three-profile solution shows a significant improvement in model fit as evidenced by a significant reduction in AIC (51750.54), BIC (51845.58) and adjusted BIC (51801.09) compared to the two-profile model. Besides, the entropy value of the three-profile solution (0.883) is also high, indicating high classification accuracy. The Lo-Mendell-Rubin (LMR) test also yielded a significant result for the three-profile solution ($p < .05$), further supporting the addition of a third profile. Although the four and five-profile solutions have lower AIC, BIC, and ABIC values, the entropy value (0.883) drops significantly, indicating that the classification is less accurate. In addition, the LMR test results indicated that there was no further support for the addition of the fourth profile ($p > .05$). The 3-profile solution provides a balanced and meaningful structure and was selected as the most appropriate model for further analysis. After the 3-profile model was selected as the optimal solution, a closer look at this model was taken.

The data presented in Table 4 highlight the means for each profile across response time, number of actions, and self-report items. Figure 1 also shows the average standardized scores for three variables across different profiles.

The ANOVA results indicated that there were statistically significant differences between the profiles for all three variables ($p < .05$). In post-hoc analyses, the Tukey test was performed to examine the differences between profiles. Tukey test results indicated that all profiles were significantly different on all variables (RT, NA and SRE, $p < .05$).

The first profile (Profile 1) consists of 5431 students representing 82.79% of the sample and is characterized by a low number of actions within a short time period, i.e. they didn't put a high amount of effort, but they have the highest level of SRE among the three profiles ($p < .05$). They have lower RT and NA scores than Profile 2, but they are higher than Profile 3. Profile 2 consists of 472 students (7.20%) who have the highest mean response time ($p < .05$) and number of actions ($p < .05$), indicating that the test-takers exerted a high level of effort and demonstrated a low level of disengagement. Although they rated their effort lower than in the first profile ($p < .05$), it is at a moderate level and much higher than in Profile 3. Profile 3 ($n = 657$; 10.01%) had the lowest RT, NA, and SRE scores, all of which were statistically significantly different from the other profiles. This profile had the characteristics of disengaged responders and was labeled "Disengaged". Although Profile 2 had a slightly lower SRE than Profile 1, it has the highest RT and NA scores, and this pattern indicates the characteristics of "highly engaged" responders. Profile 1, with the largest number of students, had scores very close to the mean. It shows signs of engagement, but the level of engagement is lower than Profile 2, which results in the label of "Moderately-Engaged". Finally, the three-profile solution clearly distinguishes between engaged and disengaged individuals. It proved effective in differentiating between engaged and disengaged individuals. P3 is the group with the highest level of disengagement, while P2 has the highest level of engagement and P1 has moderately engaged individuals.

**Table 3.**

*LPA models fit indices with different latent profiles*

| | Two-Profile | Three-Profile | Four-Profile | Five-Profile |
|---|---|---|---|---|
| Fit Statistics | | | | |
| AIC | 52853.5 | 51750.54 | 51205.58 | 50719.29 |
| BIC | 52921.39 | 51845.58 | 51327.77 | 50868.65 |
| ABIC | 52889.61 | 51801.09 | 51270.57 | 50798.74 |
| Entropy | 0.929 | 0.883 | 0.797 | 0.806 |
| LMR (p) | 2171.753 (.00) | 1080.232 (.016) | 537.668 (.379) | 480.611 (.035) |
| Profile size (%) | | | | |
| P1 | 688 (10.49%) | 5431 (82.79%) | 4424 (67.44%) | 4453 (67.88%) |
| P2 | 5872 (89.51%) | 472 (7.20%) | 107(1.63%) | 558 (8.51%) |
| P3 | | 657 (10.01%) | 640 (9.76 %) | 1145 (17.45%) |
| P4 | | | 1389 (21.17%) | 364 (5.55%) |
| P5 | | | | 40 (0.61%) |

**Table 4.**

*Average Standardized Scores for Three Profiles*

| | P1 | P2 | P3 |
|---|---|---|---|
| Response Time Mean | -.037 | .813 | -.384 |
| Number of Actions Mean | -.164 | 1.996 | -.357 |
| Self-Reported Effort Item Mean | .267 | .001 | -2.308 |

**Figure 1.**
*RT, NA and SRE averages by profile*

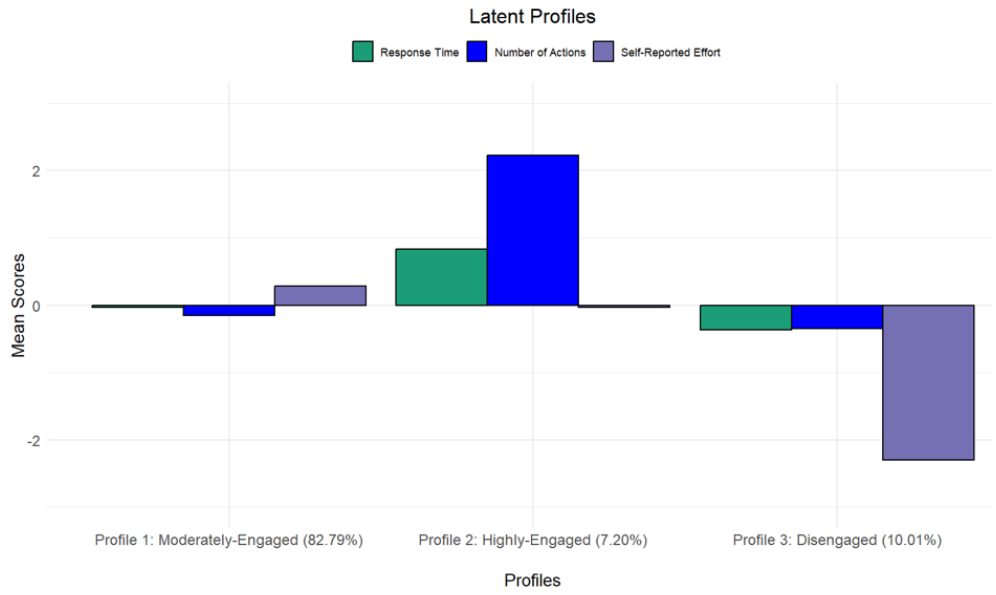



**Figure 2.**
*Profile - specific mathematics achievement means*

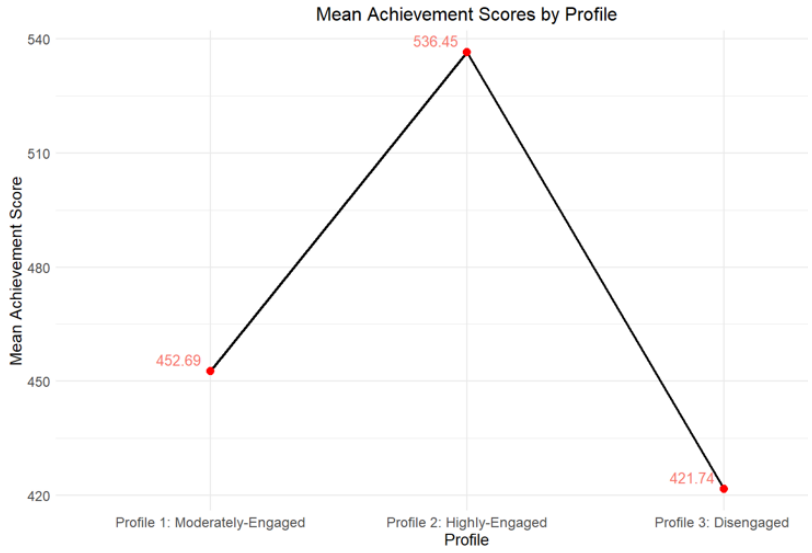



**Table 5.**
*Gender distribution at student profiles*

| Profile | n | % of Total | # of Females |
|---|---|---|---|
| Moderately-Engaged | 5431 | 82.79% | 2786 (51.30%) |
| Highly-Engaged | 472 | 7.20% | 216 (45.76%) |
| Disengaged | 657 | 10.01% | 212 (32.27%) |

N: Sample Size

While examining the mathematics achievement scores of three different profiles, it is indicated that highly engaged group (Profile 2) has the highest achievement score of 536.453. While the disengaged group has the lowest achievement score (421.74), Profile 1 has an average achievement score of 452.689. The differences were at a significant level for each group. Figure 2 shows the mean achievement scores across the different profiles.

After these interpretations, the distribution of gender was also studied in three profiles. Table 5 shows the corresponding information.

While the number of men and women in the moderately-engaged group is close, the proportion of men in the disengaged groups is almost double that of women. In the highly engaged group, the numbers of men and women are close, but tend to be dominated

by men.

**Discussion**

The purpose of the present study is to examine the test-taking disengagement behaviors of responders based on response time, number of actions, and self-reported effort data from PISA 2022 data. The results of this study provide valuable insights into student engagement and its relationship with test performance and demographic factors. Through a combination of descriptive statistics, correlation analysis, and latent profile analysis (LPA), several important conclusions emerge regarding the nature of student disengagement and its implications for educational assessment.

First, we observed that the proportion of engaged behaviors in the dataset differed significantly depending on the metrics used (RTE_5sec, RTE_10p, and NA_10p). Between response time methods, the RTE_10p method produced more conservative results compared to RTE_5sec and fewer individuals were classified as fully engaged. In the literature, item-specific threshold methods (such as normative methods), are recommended as a useful criterion to find the invalid results due to low effort (Goldhammer et. al, 2017; Wise & Ma, 2012) because they use the item characteristics too. However, it should be noted that the thresholds coinciding with 10 percent were too high and the 10 second threshold which was set as the maximum was used for all of the questions. Thus, the normative method became a common method using the 10-second threshold. On the other hand, the NA_10p method classified most students (52%) as low-engaged, suggesting that it captures a broader, potentially inflated range of disengaged behaviors. Unlike response time, where minimal time clearly signals disengagement, the number of actions (NA) may not have a straightforward relationship with cognitive effort. Certain items in the assessment may naturally require fewer actions to complete, regardless of the level of engagement or cognitive effort. On the other hand, the observed discrepancy may also stem from the threshold setting process since we have just adapted the RTE formula into the number of actions. Therefore, the method has some limitations, as the threshold used may not be universally applicable and reliable. Further research is necessary to refine these criteria. Readers should be mindful of these limitations and interpret the results with caution. These factors highlight the complexity of using the number of actions as a sole indicator of engagement and the need for careful selection of thresholds and the potential benefits of combining multiple metrics for a more comprehensive understanding of student engagement.

The correlations between response time (RT), number of actions (NA), self-reported effort (SRE), and mathematics achievement reveal some important patterns in test-taking disengagement. In particular, number of actions had the strongest correlation with performance ($r$ = .43), suggesting that higher levels of interaction with the test are positively associated with performance. The relationship between response time and achievement was also positive and at a moderate level ($r$ = .40), as observed in recent literature (Eichmann et al., 2020; Kuhfeld & Soland, 2020; Wise&Kong, 2005). However, the correlation was relatively weaker compared to the NA, in line with the findings of Csányi & Molnár (2023). Conversely, self-reported effort has the weakest correlation with performance ($r$ = .02), highlighting a potential gap between perceived and actual effort. The moderate correlation between RT and NA ($r$ = .37) suggests that students who spend more time on tasks also tend to perform more actions, which is consistent with higher engagement. These results indicate that log data based measures such as RT and NA are more reliable indicators of engagement and effort than self-reported measures. Previous studies have consistently shown that test-taking effort, especially when assessed using response time effort, has a stronger correlation with performance than self-reported effort (Rios et al., 2014; Silm et al., 2020 Wise & Kong, 2005).

The latent profile analysis identified three distinct engagement profiles: Moderately Engaged (Profile 1), Highly Engaged (Profile 2), and Disengaged (Profile 3). The Moderately-Engaged group, which comprised the majority (82.79% of the sample), was characterized by average RT and NA scores but the highest self-reported effort. The Highly Engaged group (7.20%) has the highest RT and NA scores, indicating sustained effort on the task, despite slightly lower self-reported effort than the Moderately Engaged group. The Disengaged group (10.01%) has the lowest RT, NA, and SRE scores, highlighting their lack of effort and interaction during the test. Math achievement scores varied significantly across the engagement profiles, further validating the LPA results. These performance differences underscore the critical role of engagement in academic success and suggest that targeted interventions to increase engagement could significantly improve achievement.

An analysis of the gender distribution also reveals interesting trends. While the 'Moderately Engaged' group includes almost equal numbers of men and women, the 'Disengaged' group is more prevalent among men, with almost twice as many men as women in the Disengaged profile. Conversely, the Highly Engaged group shows a slight male predominance, although the difference is not as great. These patterns suggest potential gender differences in engagement behaviors, in line with the findings in the literature (Buchholz et al., 2022; DeMars et al., 2013; Wise et al., 2010) which warrant further investigation

to understand the underlying causes and address inequalities.

In conclusion, this study highlights the multifaceted nature of student engagement and its critical influence on academic outcomes. This study provides an important step towards a better understanding of students' behavior and effort during the exam process. The findings obtained with the LPA method suggest that test-taking effort can be modeled in different profiles and that these profiles should be taken into account in exam design and assessments. Rather than focusing solely on exam outcomes, educational systems should devise more equitable and efficient assessment approaches by considering students' effort and motivation throughout the examination process. Policymakers and educators should consider using multiple engagement metrics, such as response time and number of actions, alongside measures of motivation, to create a more holistic picture of student performance. By addressing both effort and motivation across diverse contexts, education systems can better support student learning and equity worldwide.

Future research should explore alternative threshold settings for the number of action and focus on refining response time and action-based metrics to better identify disengagement, particularly through the use of item-specific thresholds for both number of action and response time which could provide more accurate and context-sensitive measures of engagement. This would help refine our understanding of how cognitive engagement is reflected across different types of test items and lead to more valid and reliable classifications of engagement. A crucial dimension to explore further is the role of motivational factors in engagement behaviors. Investigating these factors across different demographic groups, including gender, socio-economic status, and cultural contexts, can provide insights into disengagement and help develop targeted interventions. Another area of interest is cross-national comparisons of engagement behavior. Our study was limited to the Turkish sample, but examining how students' engagement and motivational factors differ across countries could provide a broader perspective on how educational systems, cultural values and socio-economic conditions shape test-taking behavior. By identifying best practices in countries with higher levels of participation, such analyses can provide actionable strategies for improvement in other regions.

**References**


Akyol, P., Krishna, K. & Wang, J. (2021) Taking PISA seriously: How accurate are low-stakes exams?. *Journal of Labor Research*, 42, 184–243 (2021). https://doi.org/10.1007/s12122-021-09317-8

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, 16, 441-462. https://doi.org/10.1007/BF03173192

Buchholz, J. (2022), "Are students trying hard to succeed in PISA?", *PISA in Focus*, No. 119, OECD Publishing, Paris, https://doi.org/10.1787/16c159b2-en

Buchholz, J. , Cignetti, M. & Piacentini, M. (2022). Developing measures of engagement in PISA. *OECD Education Working Papers* (279). https://dx.doi.org/10.1787/2d9a73ca-en

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). Harper & Row.

Csányi, R. & Molnár, G. (2023). How do test-takers rate their effort? A comparative analysis of self-report and log file data. *Learning and Individual Differences*, *106*(2023). https://doi.org/10.1016/j.lindif.2023.102340

DeMars, C. E., Bashkov, B. M. & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, 8, 69-82.

Eichmann, B., Greiff, S., Naumann, J., Brandhuber, L., & Goldhammer, F. (2020). Exploring behavioural patterns during complex problem-solving. *Journal of Computer Assisted Learning*, *36*(6), 933–956. https://doi.org/10.1111/jcal.12451

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345–356. https://doi.org/10.1080/0969594X.2010.516569

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS advanced. *Applied Measurement in Education, 27*(1), 31–45. https://doi.org/10.1080/08957347.2013.853070

Ferguson, S. L., G. Moore, E. W., & Hull, D. M. (2020). Finding latent groups in observed data: A primer on latent profile analysis in Mplus for applied researchers. *International Journal of Behavioral Development*, *44*(5), 458-468. https://doi.org/10.1177/0165025419881721

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, *2015*(2), 1-17. https://doi.org/10.1002/ets2.12067

Gignac, G. E., Bartulovich, A., & Salleo, E. (2019). Maximum effort may not be required for valid intelligence test score interpretations. *Intelligence*, 75, 73–84. https://doi.org/10.1016/j.intell.2019.04.007

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC. *OECD Education Working Papers*, (No. 133). https://dx.doi.org/10.1787/5jlzfl6fhxs2-en

Goldhammer, F., Martens, T. & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessment in Education*, *5*(18). https://doi.org/10.1186/s40536-017-0051-9

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Hofverberg, A., Eklöf, H., & Lindfors, M. (2022). Who makes an effort? A person-centered examination of motivation and beliefs as predictors of students' effort and performance on the PISA 2015 science assessment. *Frontiers in Education*, *6* (2021). https://doi.org/10.3389/feduc.2021.791599

Kong X. J., Wise S. L. & Bhola D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*, 606–619. doi: 10.1177/0013164406294779

Kuhfeld, M. & J. Soland (2020). Using Assessment Metadata to Quantify the Impact of Test Disengagement on Estimates of Educational Effectiveness, *Journal of Research on Educational Effectiveness*, *13*(1), 147-175, https://doi.org/10.1080/19345747.2019.1636437

Li, C. H. (2015). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

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

Messick, S. (1984). The nature of cognitive styles: problems and promise in educational practice. *Educational Psychologist, 19*, 59-74. https://doi.org/10.1080/00461528409529283

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. doi:10.2307/1175249

Morgan, G. B. (2015). Mixed mode latent class analysis: An examination of fit index performance for classification. *Structural Equation Modeling*, *22*(1), 76–86. https://doi.org/10.1080/10705511.2014.935751

Muthén, L. K., & Muthén, B. O. (2014). *Mplus: Statistical Analysis with Latent Variables: User's Guide* (Version 7).

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535–569. https://doi.org/10.1080/10705510701575396

OECD (2016). *Low-performing students: Why they fall behind and how to help them succeed*. PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264250246-en

OECD (2024). *PISA 2022 technical report*. PISA, OECD Publishing. https://doi.org/10.1787/01820d6d-en

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 Assessment in Education*, *9* (10),. https://doi.org/10.1186/s40536-021-00104-6

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rios J. A., Liu, O. L. & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research*, *2014* (161). https://doi.org/10.1002/ir.20068

Sahin, F., & Colvin, K.F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-scale Assessment in Education*, *8*(5). https://doi.org/10.1186/s40536-020-00082-1

Soland, J., Wise, S. L., & Gao, L. (2019). Identifying disengaged survey responses: New evidence using response time metadata. *Applied Measurement in Education*, *32*(2), 151-165. https://doi.org/10.1080/08957347.2019.1577244

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior, 120,* 103445. https://doi.org/10.1016/j.jvb.2020.103445

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge University Press.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretations, and implications. *Educational Measurement Issues & Practice*, *36*(4), 52–61. https://doi.org/10.1111/emip.12165

Wise, S. L. (2020). The impact of test-taking disengagement on item content representation. *Applied Measurement in Education*, *33*(2), 83–94. https://doi.org/10.1080/08957347.2020.1732386

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1-17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, *53*(1), 86-105. https://doi.org/10.1111/jedm.12102

Wise, S. L., & Kingsbury, G. G. (2022). Performance decline as an indicator of generalized test-taking disengagement. *Applied Measurement in Education*, *35*(4), 272–286. https://doi.org/10.1080/08957347.2022.2155651

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010). *An investigation of the relationship between time of testing and test-taking effort*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver.

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method*. In annual meeting of the National Council on Measurement in Education, Vancouver, Canada (pp. 163-183).

Zamarro, G., Hitt, C., & Mendez, I. (2019). When students don't care: Reexamining international differences in achievement and student effort. *Journal of Human Capital*, *13*(4), 519-552. https://doi.org/10.1086/705799