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_β, 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

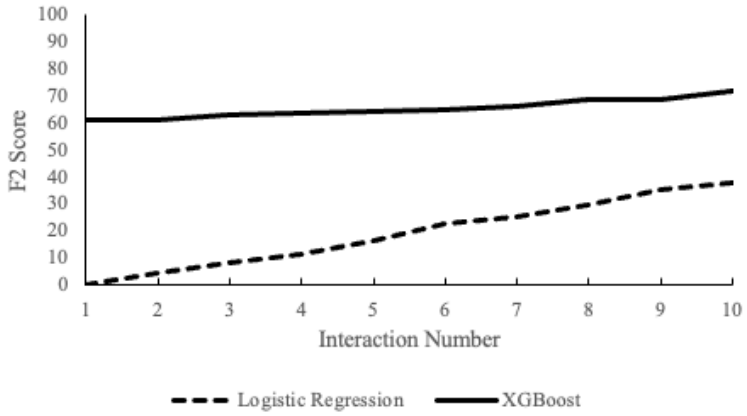
Interaction number #	TP	TN	FP	FN	F2 Score	Accuracy	Recall
1	1230	530	3750	30	61.48	31.79	98.01
2	1020	2060	2220	240	61.47	55.58	80.88
3	1050	2050	2240	200	63.36	55.97	83.90
4	1040	2180	2100	210	63.93	58.27	83.19
5	1010	2470	1810	250	64.34	62.85	80.40
6	990	2670	1620	270	64.80	66.01	78.73
7	1020	2690	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.

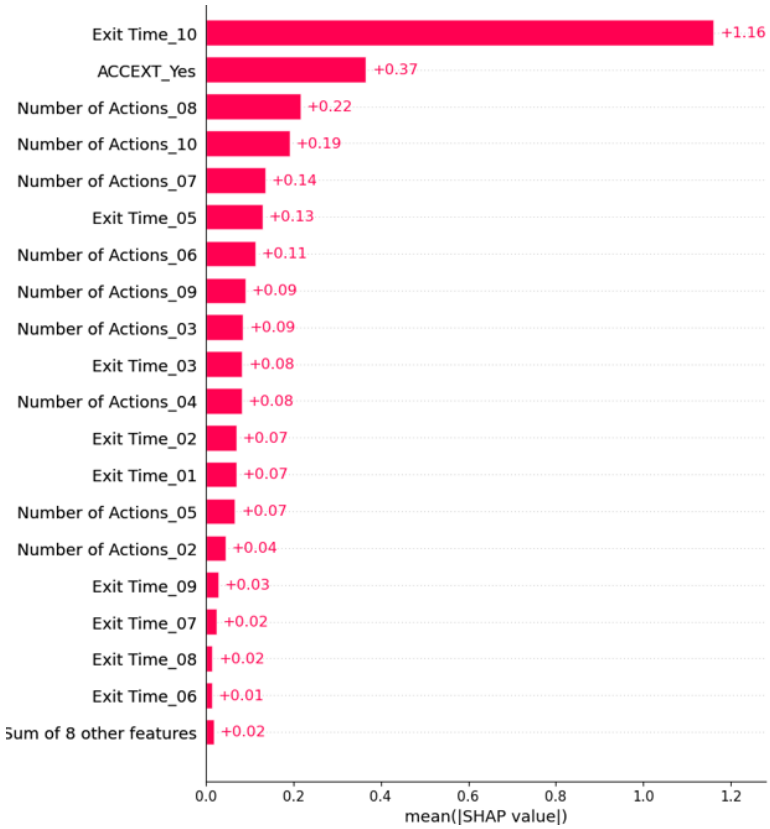
During the interaction with the first item, the model demonstrated a high recall rate of 98.01%, successfully identifying 1,230 TPs. It achieved an accuracy of 31.79% and an F2 score of 61.48, indicating a strong ability to

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.

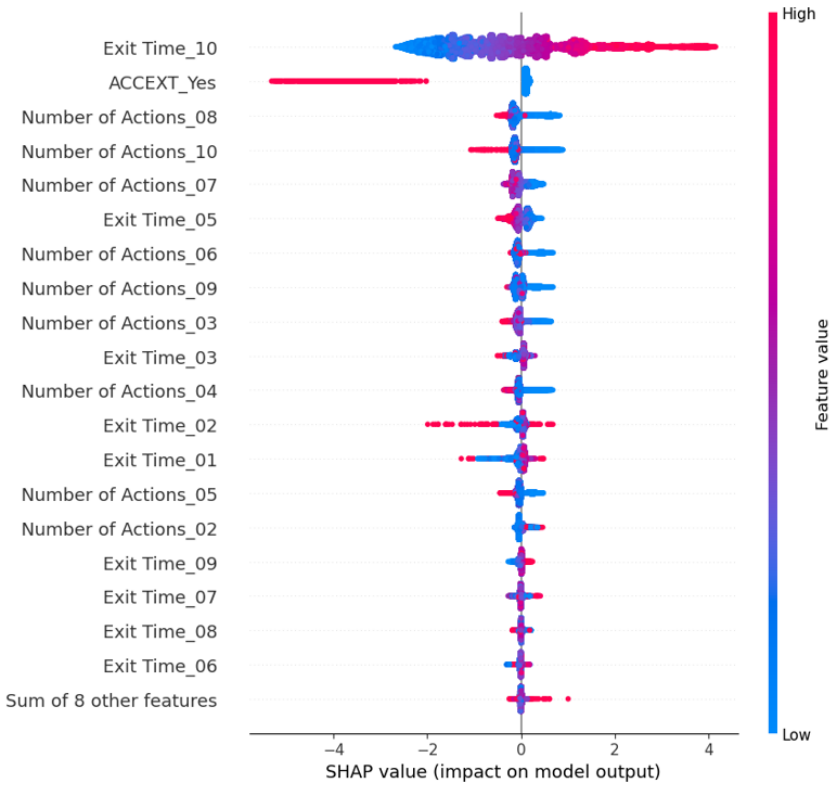
The SHAP (Shapley Additive exPlanations) values, a game-theoretic approach designed to explain the output of machine learning models (Lundberg, et al., 2020), were used in interpreting the influence of model features on predictions. For the 10th model, we examined the SHAP values through various visualizations. 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

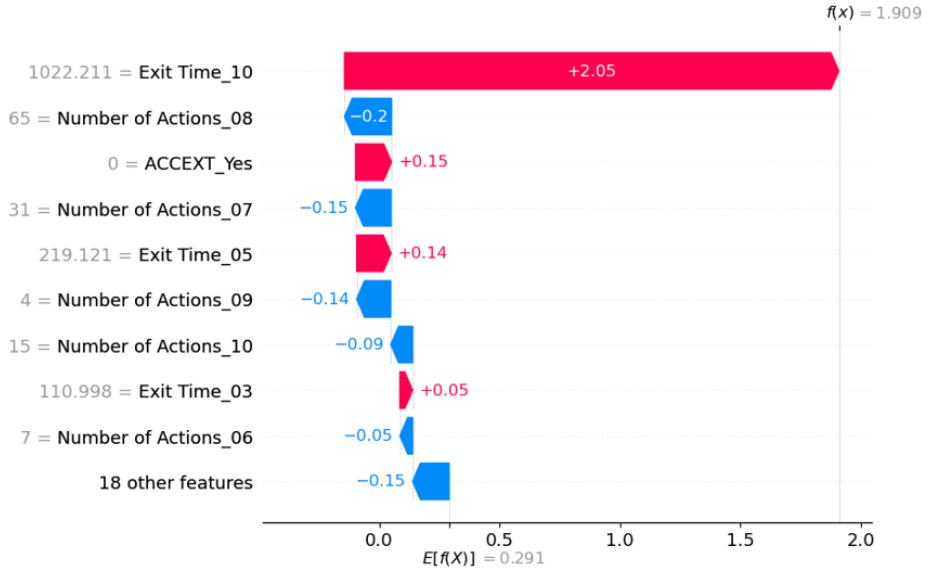
Figure 3.

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



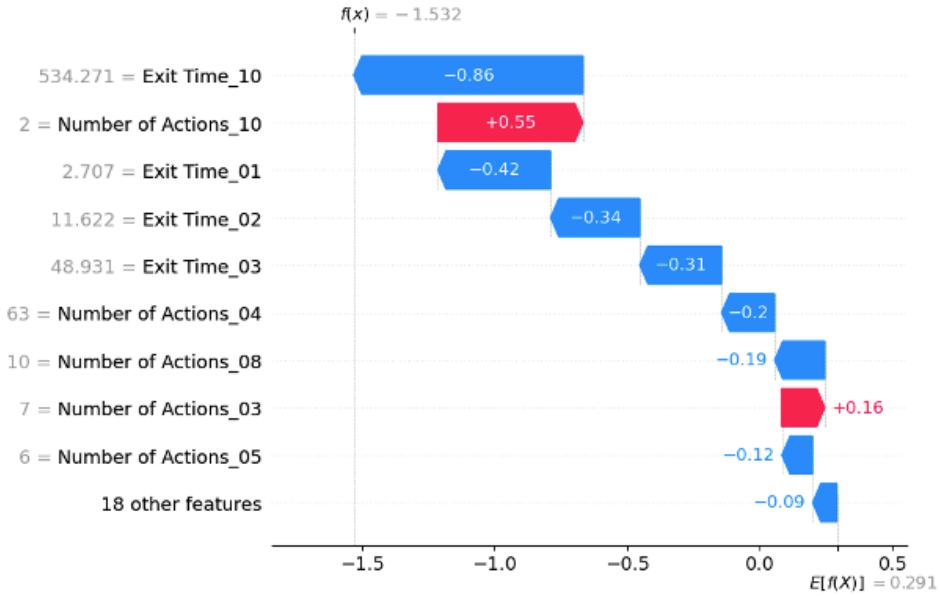
Data source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade 8 Mathematics Assessment.

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)

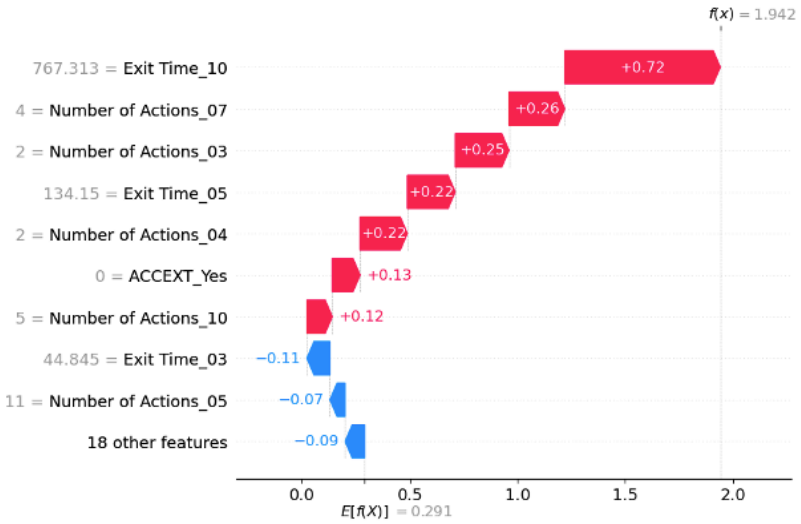


Note. Blue bars represent features that lower the prediction, such as the exit time for the interaction with the 10th task.
Data source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Grade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

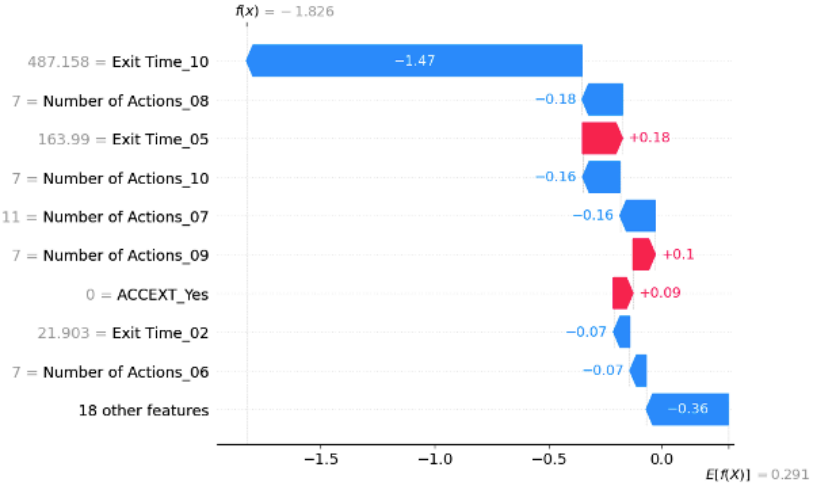
Notably, the exit time for the 10th item also plays a critical role in the misclassification of cases, influencing both false positives and false negatives. This is evident in Figures 5 and 6, where the impact of shorter or longer exit times, respectively, steers the model's predictions, affecting its 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