ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(Special issue); 282-301

Human-Centered AI for Discovering Student Engagement Profiles on Large-Scale Educational Assessments

Hongwen GUO* Matthew JOHNSON** Luis SALDIVIA***
Michelle WORTHINGTON**** Kadriye ERCIKAN*****

Abstract

Large-scale assessments play a key role in education: they provide insights for educators and stakeholders about what students know and are able to do, which can inform educational policies and interventions. Besides overall performance scores and subscores, educators need to know how and why students performed at certain proficiency levels to improve learning. Process/log data contain nuanced information about how students engaged with and acted on tasks in an assessment, which hold promise of contextualizing a performance score. However, one isolated action event observed in process data may be open to multiple interpretations. To address this challenge, in the current study, use of multi-source data (performance and process) was proposed to integrate sequential process data with response data to create engagement profiles to better reflect students' test-taking processes and knowledge states. Most importantly, AI algorithms were used to assist and amplify human expertise in the creation of students' engagement profiles, so that the information extraction from the multi-source data can be scaled up to enhance the value of large-scale assessments in teaching and learning. Various machine learning techniques were leveraged to develop the general framework of the human-centered AI (HAI) approach to help human experts efficiently and effectively make sense of the multi-source data. Using a mathematics item block from the National Assessment of Educational Progress (NAEP) for illustrations, data from over 14,000 students resulted in ten preliminary profiles, more than half of which were associated with low performing students. Such HAI approaches and data insights are expected to generate rich and meaningful feedback for educators and stakeholders.

Keywords: Multi-source data, machine learning, human-in-the-loop, visualization, Large-scale assessment

Introduction

Large-scale assessments (LSAs) play a crucial role in assessing and improving the quality of education at state-, national-, and international-levels. These measures inform educators and stakeholders on what students know and can do, so that they can prepare for education policies and interventions in teaching and learning (Gordon, 2020; Pellegrino, 2020). These assessments may also help guide resource allocation in education (NAGB, 2024b). However, for educators to use these large-scale assessment results in a classroom, a performance score may not be enough, particularly for low performing students. Educators need to know how and why these students got low scores, so that they can prepare targeted and effective interventions. In the rapidly evolving landscape of educational technologies, many LSAs are administered on digital platforms, where students' digital footprints (i.e., process/log data) are collected (Ercikan & Pellegrino, 2017; Ercikan et al., 2023). These process data contain nuanced information about how students solved the tasks and how they navigated through the assessment, which may reflect students' cognitive thinking processes, affective states, and test-taking strategies, holding promise of providing contextual information beyond a performance score.

* Principal Research Scientist, ETS, Princeton-New Jersey, US, hguo@ets.org, ORCID ID: 0000-0002-1751-0918
**Principal Research Director, ETS, Princeton-New Jersey, US, msjohnson@ets.org, ORCID ID: 0000-0003-3157-4165
*** Research Strategic Advisor, ETS, Princeton-New Jersey, US, lsaldivia@ets.org, ORCID ID: 0009-0007-3482-7654
**** Assessment Development Manager, ETS, Princeton-New Jersey, US, mworthington@ets.org, ORCID ID: 0009-0006-0480-3769
***** SVP of Global Research, ETS, Princeton-New Jersey, US, kercikan@ets.org, ORCID ID: 0000-0001-8056-9165

To cite this article:
Guo, H., Johnson, M., Saldivia, L., Worthington, M. & Ercikan, K. (2024). Human-Centered AI for discovering student engagement profiles on large-scale educational assessments. *Journal of Measurement and Evaluation in Education and Psychology, 15*(Special issue), 282-301. https://doi.org/10.21031/epod.1532846

Received: 14.05.2024
Accepted: 24.09.2024

As described in prior literature (Ercikan et al., 2023), the key uses of process data in assessments include score validation (Wise, 2021), assessment design improvement (Pellegrino, 2020), evidence for the targeted construct (Johnson & Liu 2022; Levy, 2020; Pohl et al., 2021), group comparison (Ercikan et al., 2020; Guo & Ercikan, 2021a, 2021b; Rios & Guo, 2020), and feedback enrichment (Guo et al., 2024; Zoanetti & Griffin, 2017). Many features have been extracted from process data, among which time on task is one of the most-commonly used features. Item response times have been shown to be significantly associated with performance on LSAs (Ercikan et al., 2020; Guo & Ercikan, 2021a, 2021b; Rios et al., 2017; Wise, 2017, 2021). To solve an item, an appropriate amount of time needs to be spent in understanding the question and working towards its solution. A hard item usually takes a longer time to solve, and an easy item a shorter time. On LSAs, certain rapid responding behaviors associated with guessing are often observed, which may compromise score validity. However, such behavior can be associated with varied factors, such as low-test motivation, specific test-taking strategies (e.g., skipping hard items), and speededness because of time pressure. A rapid response may also be observed from a student simply because of their high proficiency and efficiency. Without context, it is challenging to explain why students rapidly respond to an item on a test. Other process data features face similar challenges in interpretation as well, since one isolated behavior observed during the test-taking process may be open to multiple interpretations (Ercikan et al., 2023; Greiff et al., 2016; Guo et al., 2024).

To address these challenges, the current study proposes to integrate sequential process data with response data (also called multi-source data in the current study) to create engagement profiles to better reflect students' test-taking processes for rich insights beyond their knowledge and skills in a knowledge domain. More specifically, in the current study, the multi-source data for each student (i.e., the item navigation sequence, the response time sequence associated with each item navigation, and the response score sequence corresponding to the items) are used to create preliminary profiles. These profiles would inform educators and stakeholders not only what a performance level of a student or a student group reached, but also how they worked through the assessment, which in turn would help to shed light on why they performed at this level. Such information and data evidence are particularly useful for helping low performing students.

Most importantly, given the large sizes of data students produced on LSAs, we propose to use AI/machine learning algorithms to assist and amplify human expertise in the creation of engagement profiles, so that the information extraction from these multi-source data can be scaled up to enhance the impact of large-scale assessments in teaching and learning. Therefore, one goal of the current study is to propose a general framework that uses AI to augment human experts in uncovering data insights and expediting the development of student profiles on a large scale. The engagement profiles created in the study may reflect students' navigation processes, affective states, and test-taking strategies, among others. Note that the engagement profiles use students' sequential information in the response and process (i.e., timing and navigation) data when they interacted with the assessment platform, which provide richer context than the commonly used engagement indices (such as the response time effort proposed by Wise and colleagues, 2005, 2017, 2021), but requires more intensive computational demands.

To create engagement profiles, our research questions are:

- RQ-1. What are the considerations in data preprocessing? This includes the creation of meaningful and explainable process data features and data visualization to assist human experts.
- RQ-2. How to start from scratch for human experts to discover engagement profiles? Since the proposed engagement profiles are novel, it is necessary to discover them from data. Given the expected large sizes of students on LSAs and the volume of process data students produced, we need an efficient approach to select a manageable and informative sample of students' data for human experts to examine and discover the initial engagement profiles.

- RQ-3. How to scale up the engagement profile creation? That is, how to combine the unique strengths of AI algorithms and human knowledge, thereby improving overall performance and productivity in the profile creation for all students.

The paper is structured as follows. In the Method section, we introduce the large-scale assessment, response data, and process data used in the study. Then we present the proposed Human-centered AI (HAI) framework for data analysis and profile creation. Three major steps in the HAI architecture are described in detail to show how human knowledge plays a crucial role in the profile creation, how to leverage AI algorithms (such as machine learning, deep learning, and active learning methods) to enhance data analysis and pattern identification, and how to combine AI power and human expertise to create the profiles. In the Results section, we present results obtained from each of the three major steps in the HAI architecture. In the last section, we discuss the potential uses of the engagement profiles, the implications and significance of the HAI general framework, and limitations of our current work.

Methods

Research Design

HAI approaches have been strongly recommended in education to make decisions based on established, modern learning principles, wisdom of educational practitioners, and human knowledge in the educational assessment community (Baker, 2021; Guo et al., 2024; Miao et al., 2021). In this study, the application of HAI is intended to assist and amplify (rather than displace) human expertise in understanding students' knowledge, skills, and abilities (KSAs) "beyond a sole focus on students' core academic performance measured by large-scale assessments, to support students and teachers with actionable feedback that nurtures the broader skills students need to succeed and thrive" (Office of Educational Technology, 2023).

In this study, data from the National Assessment of Educational Progress (NAEP) Grade 8 Mathematics assessment were used for illustration. NAEP provides important information about student academic achievement and learning experiences in various subjects and has provided meaningful results to improve education policy and practice in the US.

The NAEP mathematics assessment was first administered digitally in 2017. This digital administration allowed for the collection of process data, including information on how long students spent on the assessment questions (commonly referred to as timing data), how they navigated through items, and how students used onscreen assistive digital tools to develop their responses. NAEP also releases samples of process data. Interested researchers can consult their website for more information (NAGB, 2020).

Five broad content areas in the NAEP mathematics assessment are number properties and operations; measurement; geometry; data analysis, statistics and probability; and algebra, which are measured using a variety of item types including selected responses (e.g., single-and multiple-selection multiple choice, and matching), and short or extended constructed response (CR). Items are also classified on three levels of cognitive complexity (Low, moderate, and high), based on the items' demands on students' thinking process (NAGB, 2024a).

Response Data

For this study, we selected one item block in the 2022 NAEP 8th Grade Math assessment, because it contained many publicly released items. Detailed information on content of the released items and scoring rules can be found on the National Center of Educational Statistics (NCES) website (NCES, 2022). This item block has 14 items, and students can have 30 minutes to work on it (refer to Table 1). In the "Item" column of Table 1, items with * are released items. In the "Skill" column, Data stands for Data Analysis, Statistics, and Probability; Number stands for Number Properties and Operations. In the "Item Type" column, SR stands for selected response, and CR stands for short or extended constructed response.

Table 1.
Item information

Item	Skill	Item Type	Max Score	Item Difficulty
1	E) Algebra	SR	1	Very Easy
2*	D) Data	CR	2	Medium
3	E) Algebra	SR	1	Very Easy
4	B) Measurement	SR	1	Very Easy
5*	D) Data	SR	1	Easy
6*	E) Algebra	CR	1	Easy
7	C) Geometry	SR	2	Medium
8	A) Number	SR	1	Easy
9*	E) Algebra	SR	1	Hard
10*	C) Geometry	SR	2	Easy
11	A) Number	MS	2	Medium
12*	E) Algebra	SR	1	Easy
13*	C) Geometry	CR	4	Hard
14*	B) Measurement	SR	1	Medium

The maximum item score varies from 1 point to 4 points, as shown in the "Max score" column in Table 1. The total maximum raw score on the block is 21 points. For example, Item 13 is a CR item, with a maximum score of 4. A student can get a score of 0 for completely incorrect responses, a credit of 1, 2, or 3 for partial correct responses, or a score of 4 for full credit.

Process Data

NAEP digitally based assessments offer a testing environment that makes it possible to record students' interaction with the digital platform when students solve the tasks. Figure 1 shows a screenshot of the testing environment of one released item (NAGB,2020).

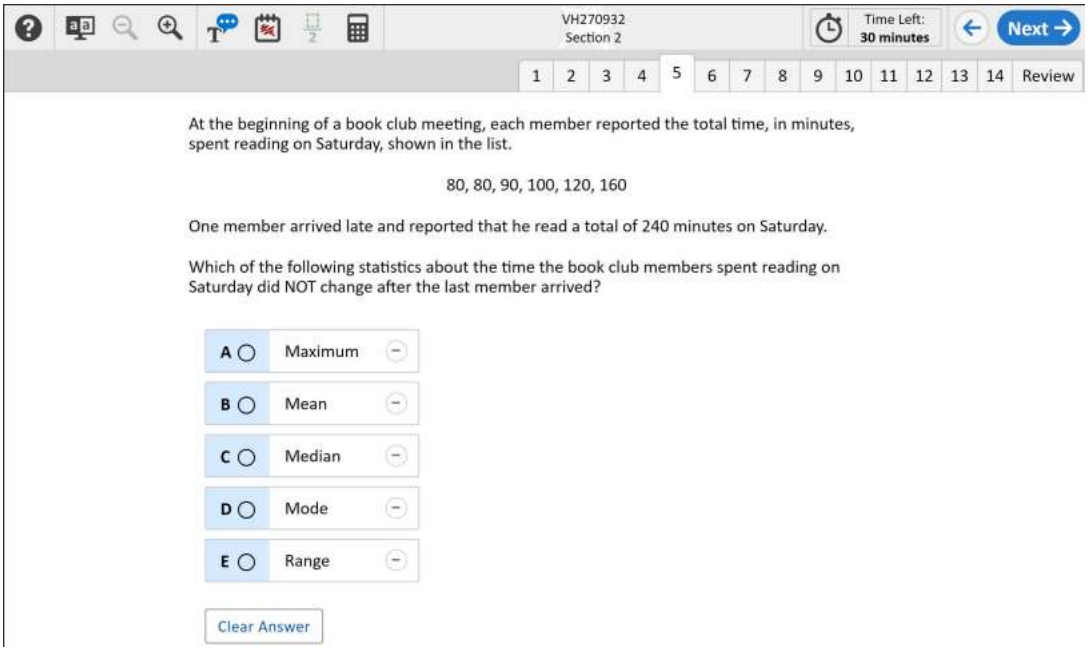
Starting from the upper left corner of the screen in Figure 1, the digital tools include help (a question mark), color contrast and theme change, zoom-in/out, text-to-speech, scratch work, equation editor, calculator (note that the studied item block allows the use of a calculator). On the upper left corner of the screen, students can monitor their session time (a clock icon), check their progress on the items, and move forward or backward of the pages/items by clicking on the item tags or using the 'Next' button. The 'Review' button allows students to get an overview of which items they had responded to and which they had not. Students' interactions with the testing environment were logged and collected to produce process data.

The process data contain logs of response processes collected from each student, such as item response time, use of the digital tools, and which items students were working on and for how long. Because of space limitations, please refer to NAGB (2020) for detailed information on the process data variables.

After removing students with irregular response time and abnormal completion on the selected item block, the sample size in the study is N=14,008.

Figure 1.

The NAEP testing environment in the 2022 Math Assessment (using one released as an illustration).



Data Analysis and Procedures

The proposed general framework of the human-centered AI (HAI) architecture (refer to Figure 2) attempts to amplify human knowledge, maximize AI power, and minimize redundancy of human labor, so that the data can be effectively and efficiently annotated to address the big data challenges.

There are three major steps in the architecture in Figure 2, each of which relies on human knowledge and decisions.

Step-1 (data preprocessing & feature engineering) includes data cleaning and feature engineering. This process is informed by insights gleaned from prior research, literature, and experiences on process data and test-taking behaviors on educational assessments.

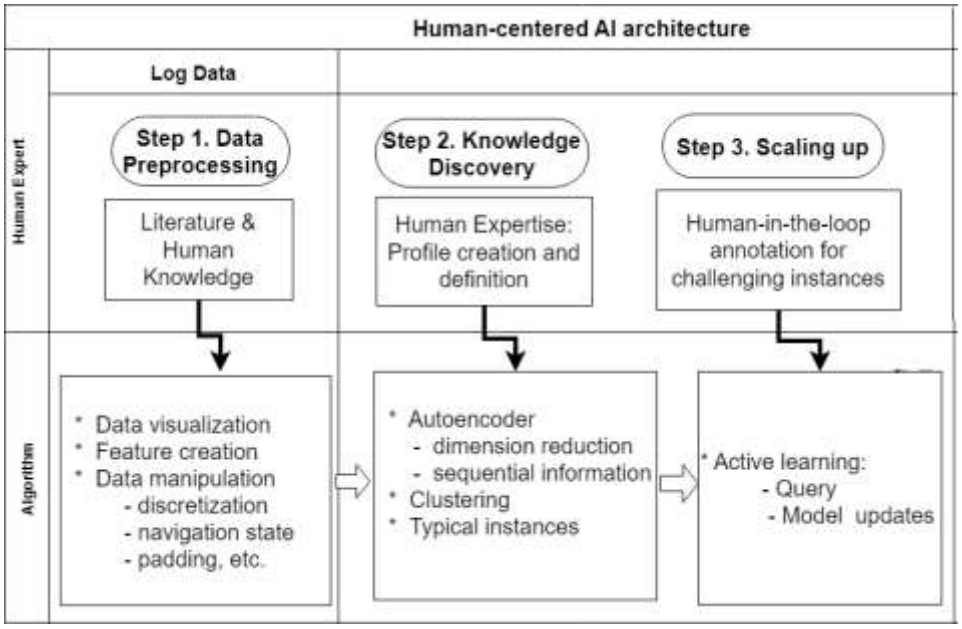
Step-2 (Knowledge Discovery) contains two parts: Part 1 uses an autoencoder (a self-supervised deep learning model) to compress the input sequential data (item responses, response times, and item navigation states) into a low-dimensional space (also called the latent space or the code space). Part 2 uses a clustering method to select typical data patterns for human experts to discover engagement profiles.

Step-3 (Scaling up) uses the active learning method to amply human experts' knowledge to unlabeled data.

A similar HAI architecture was applied in a recent study that investigated digital assistive tool uses, response times, and performance on the assessment platform (Guo et al., 2024). In the following paragraphs, we provide more details for each step.

Figure 2.

The general framework of the proposed HAI architecture



Step-1. Data Preprocessing and Feature Engineering

One of the prominent features extracted from process data is item response time. As discussed in the introduction, a rapid or prolonged response time may imply unexpected behaviors on the assessment. A rapid response is likely to reflect random guessing, which adds noise to response data and does not reflect students' true knowledge and skills (Guo & Ercikan, 2021a, 2021b; Guo et al., 2017; Rios & Guo, 2020; Wise, 2021). Therefore, we created six-time categories (refer to Table 2 below) to help to interpret the meaning of the item response time, in terms of whether a student spent reasonable time on an item (Guo & Ercikan, 2021a).

Table 2.

Definition of Time Categories and Their Possible Implications

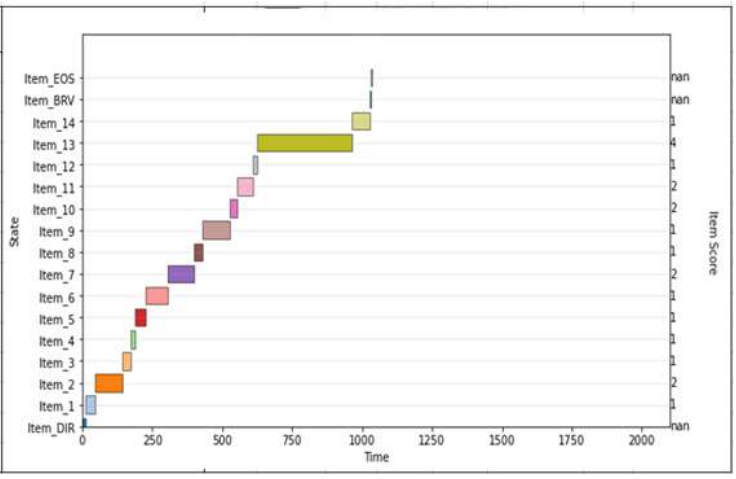
Time Category	Definition	Implication
0	$T = 0$	Student did not work on the studied item.
1	$0 < T \leq$ Threshold*	Rapid responding (likely associated with random guessing)
2	Threshold* $\leq T < Q_1$	Student may spend sufficient time (but still low).
3	$Q_1 < T \leq Q_3$	Student spent sufficient time (in the middle quartiles).
4	$Q_3 < T \leq 95^{th}$ percentile	Student spent sufficient time (in the upper quartiles).
5	$T > 95^{th}$ percentile	Student spent prolonged time (likely facing challenges)

Notes. *In Column 2,* T *stands for item response time of the studied student on the studied item. The threshold* of response time for each item is derived using the hybrid method in Guo & Ercikan (2021a) to flag response times that indicate rapid responding behaviors. The quartiles* Q_1, Q_2, Q_3 *and 95th percentile are determined by the item response time distribution of the studied item for the sample (N = 14,008).*

Figure 3.

Data visualization for one instance

Item	Time	Score	Time Category	Visit	Skill	Item Type
1	34	1	2	1	Algebra	MS
2	99	2	2	1	Stats	CR
3	30	1	1	1	Algebra	MS
4	16	1	1	1	Measurement	MS
5	37	1	2	1	Algebra	MS
6	79	1	2	1	Stats	CR
7	93	2	3	1	Geometry	MS
8	30	1	1	1	Stats	MS
9	96	1	2	1	Algebra	MS
10	26	2	2	1	Geometry	MS
11	54	2	3	1	Number	MS
12	15	1	1	1	Algebra	MS
13	357	4	4	1	Geometry	CR
14	63	1	3	1	Measurement	MS



Notes. *The student had a total score of 21 out of 21, a total time of 1029 out of 1800 seconds and a total number of visit states of 17. The item-level summary information (item score, item response time, response time category, and number of item visits) is presented for the student in the table on the left-hand side. The student's navigation pattern is presented in the plot on the right-hand side. This student visited items linearly one item at a time and was in the highest performing profile.*

To illustrate, the extracted data for each student are presented at three levels: block level, item level, and granular navigation level (refer to Figure 3 as an example for one student). In Figure 3, the item block level summary is provided in the caption. For this student, the total score received is 21 out of the maximum 21 points; the total time spent is 1029 out of the maximum 1800 seconds; and the total number of visit/navigation states, which is 17, including reading direction (Item_DIR) at the beginning of the session, reviewing the block (Item_BRV) at the end of the session, and ending the session (Item_EOS). The navigation plot on the right side of the figure is a visualization of three sequences (navigation state, time on the state, and score received). Each colored rectangle shows the time spent on an individual navigation state. In the plot, the *x*-axis stands for the testing time, the *y*-axis on the left stands for the item state (i.e., what item the student was working on) and other navigation states, and the *y*-axis on the right stands for the item score the student obtained.

Figure 3 shows that this student worked linearly through the items by the item presentation order from Item 1 to Item 14. The table on the left-hand side of the navigation plot provides the item level information, regarding time category (refer to Table 2 for definition) on an item, total time spent on the item (in seconds), item score received, number of visits on the item, skill measured, and item type. Please refer to Figures A2 to A6 in Appendix for more examples.

In the data pre-processing step, we emphasized preserving sequential information and integration of response data and process data, because one isolated event was often open to multiple interpretations as to what generated it. For example, for a low performing student, a rapid response observed at the beginning of the assessment (refer to Figure A1 in Appendix for one example) and one observed at the middle of the assessment (refer to Figure A3 in Appendix for another example) clearly contain different meanings: the earlier rapid responding behavior is likely to be an indicator of the low test-taking motivation of the student, and the latter one is likely to be an indicator of applying a test-taking strategy of skipping a question on which the student may lack knowledge. On the other hand, for a high performing student with a perfect score, a rapid response may indicate high efficiency in solving the problems (refer to Figure 3 above for one example).

Data gathered for each student, as presented in Figure 3, ensures that the features are meaningful and interpretable for human understanding and annotation, to addresses RQ-1.

Step-2. Knowledge Discovery

In order to help human experts to start the profiling process from scratch, we used an autoencoder to compress the navigation sequences into a low-dimensional space. Autoencoders possess the ability to acquire compact representations of input data, operating in a self-supervised manner wherein data labels are absent (Geron, 2017). An effective autoencoder demonstrates proficiency in reconstructing input data autonomously upon decoding the code space. Within the autoencoder architecture, we implemented the long short term memory (LSTM) layers to maintain sequential information to capture the sequence nature of students' navigation of the item block and enhance data interpretation, particularly addressing RQ-2. Step-2 also includes a key component of knowledge discovery, for which, a clustering method (an unsupervised machine learning method) was applied to identify typical and representative instances in the big data, so that human experts could work on these typical instances to make sense of students' data, discover patterns, and define profiles. In this step, a large number of clusters was chosen on purpose to help with knowledge discovery. More specifically, in the current study, the number of clusters was 30; in each cluster, three representative instances were selected, and the visualization of each instance as in Figure 2 was presented to human experts to review and create profiles. Note that additional extreme cases (such as those with the highest/lowest score, with the longest/shortest time, and with the largest/smallest number of visits) were also presented to human experts to assist in profile creation.

Step-3. Scaling up

Based on the human labeled data, in this step, we applied an active learning approach integrated with a semi-supervised learning (AL&SSL) to predict the profiles for the unlabeled students' data (Guo et al., 2024; Rizve et al., 2021; Xie et al., 2019; Zhu et al., 2003), which addresses RQ-3.

Figure 4.

The active learning framework (Image from Radwan, 2019)

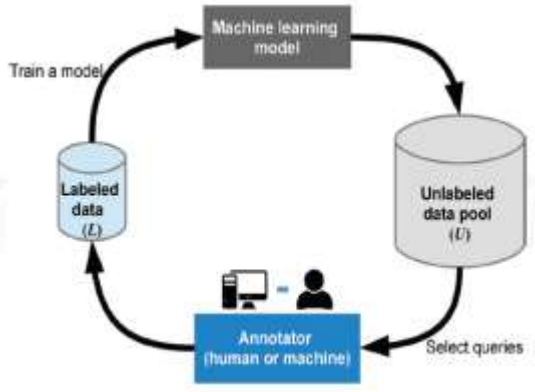
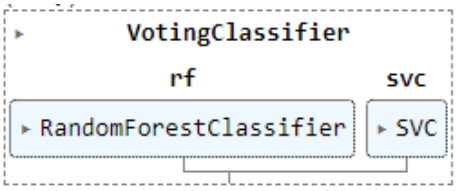


Figure 5.

The ensemble classifier



More specifically, there are two components in the AL&SSL framework (refer to Figure 4): a classifier (i.e., supervised machine learning model) and an oracle (i.e., human experts). There are four steps in one iteration in the AL&SSL framework as shown in Figure 4. We started from the "labeled data", which were obtained from Step-2 in our study. To build a "machine learning model", we used an ensemble voting classifier with a soft voting mechanism by combining a random forest classifier and a support vector machine (SVC) classifier and then trained and initialized the model with the labeled data (refer to Figure 5).

Using the trained model, we predicted pseudo labels/profiles for instances in the "unlabeled data pool". We then selected instances that were challenging to the model (i.e., instances with low confidences/probabilities for the pseudo label prediction) and asked human experts to annotate them (i.e., "human annotator" labeled data). At the same time, instances for which the model was accurate, and the prediction had high confidences, adopted the pseudo labels (i.e., "machine annotator" labeled data). The newly labelled data were then used to update both the training data and the model, and a new iteration started again. The iteration process in Figure 4 could end when all instances were labeled with satisfactory accuracy of the model and high prediction confidence of the pseudo labels.

Results

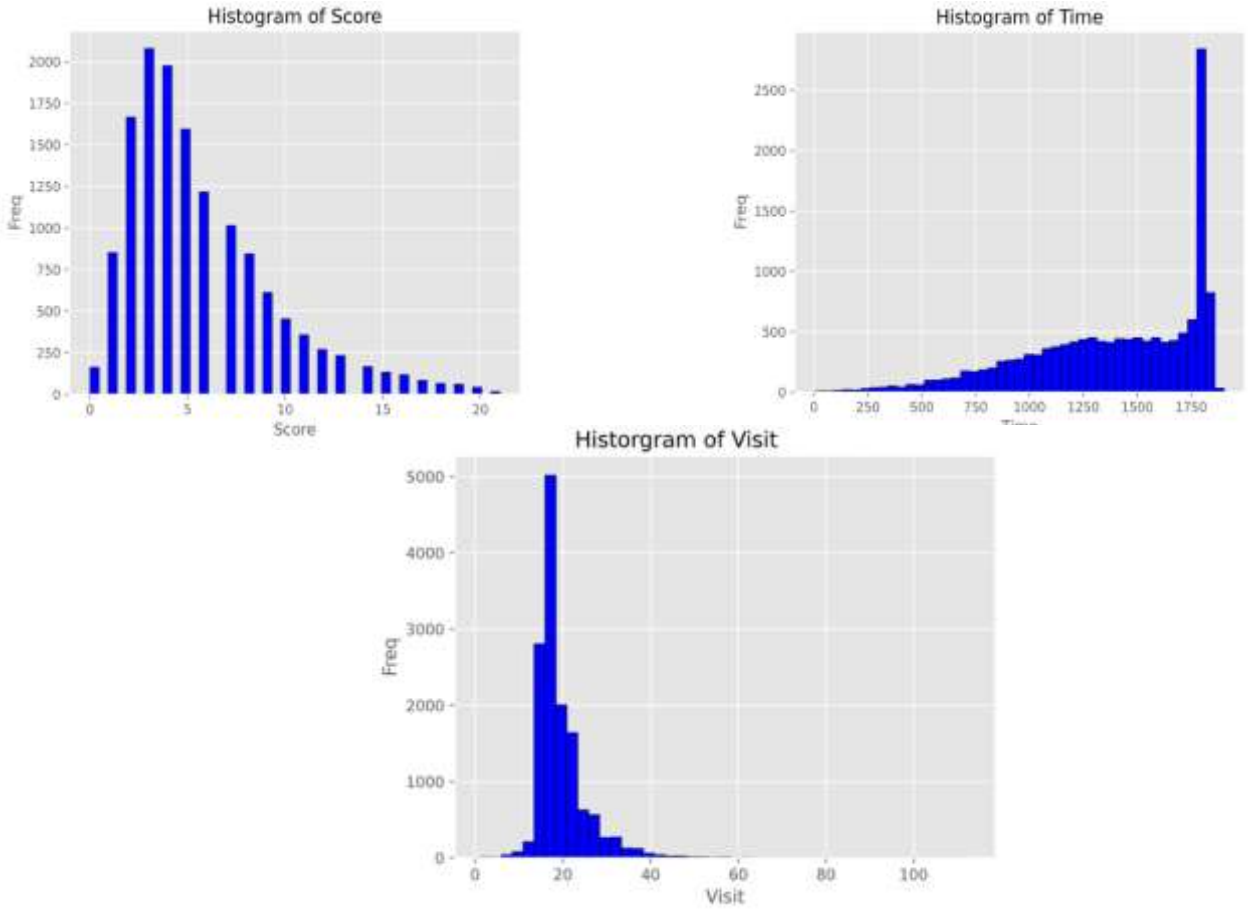
In this section, we first briefly show results from Step-1, and then we focus on the resulting engagement profiles from Step-2 and Step-3. The Python and TensorFlow libraries (Abadi et al., 2015) were used in producing the results.

Data Preprocessing Results – Step 1

As discussed in the methods section, in Step-1, we preprocessed the process data, extracted meaningful process features, and created visual presentations (as in Figure 3). In addition, Figure 6 below shows the histograms of the test-level variables (total score, total time, total number of visits) for the N=14008 students on the studied item block.

Figure 6.

The histograms of the test-level variables.



The histograms in Figure 6 show that all the test-level variables have skewed distributions: the total scores are concentrated on 3 and 4 points and have a long right tail; the total response times peak at the maximum allowed time (1800 seconds); and the total number of visits has a mode around 15 (note again that the number of items in the block is 14).

Table 3.

Item-level statistics

Item	Average Item Score	Item Difficulty	Median Time	95%tile Time	Rapid Responding Threshold	Max Score
1	0.69	0.69	43	126	13	1
2	0.99	0.50	121	310	17	2
3	0.65	0.65	65	206	38	1
4	0.50	0.50	73	233	42	1
5	0.16	0.16	69	207	42	1
6	0.25	0.25	128	336	32	1
7	0.58	0.29	90	238	25	2
8	0.17	0.17	52	144	31	1
9	0.48	0.48	148	351	45	1
10	0.27	0.14	68	207	7	2
11	0.23	0.12	40	134	20	2
12	0.41	0.41	32	94	16	1
13	0.33	0.08	195	456	24	4
14	0.10	0.10	60	202	5	1

Table 3 shows the item-level summary statistics, which shows that Item 13 (worth 4 points in total with an average item score of 0.33) is the most difficult item (difficulty is 0.08 = 0.33/4) and most time consuming (the median response time is 195 seconds); Item 1 is the easiest item (difficulty is 0.69) and second to the least time consuming (the median time is 43 seconds).

Item 12 is the least time-consuming item (the median time is 32 seconds). The 95%tile time (Time Category 5) shows that, again, Item 13 is the most time-consuming item (456 seconds), and Item 12 is the least (94 seconds). Also shown in Table 3, the thresholds for flagging rapid responses (Time Category 1) are the longest for Item 9 (45 seconds) and the shortest for Item 14 (5 seconds), respectively. For each student, the data were prepared and visualized as in Figure 3.

Knowledge Discovery Results – Step 2

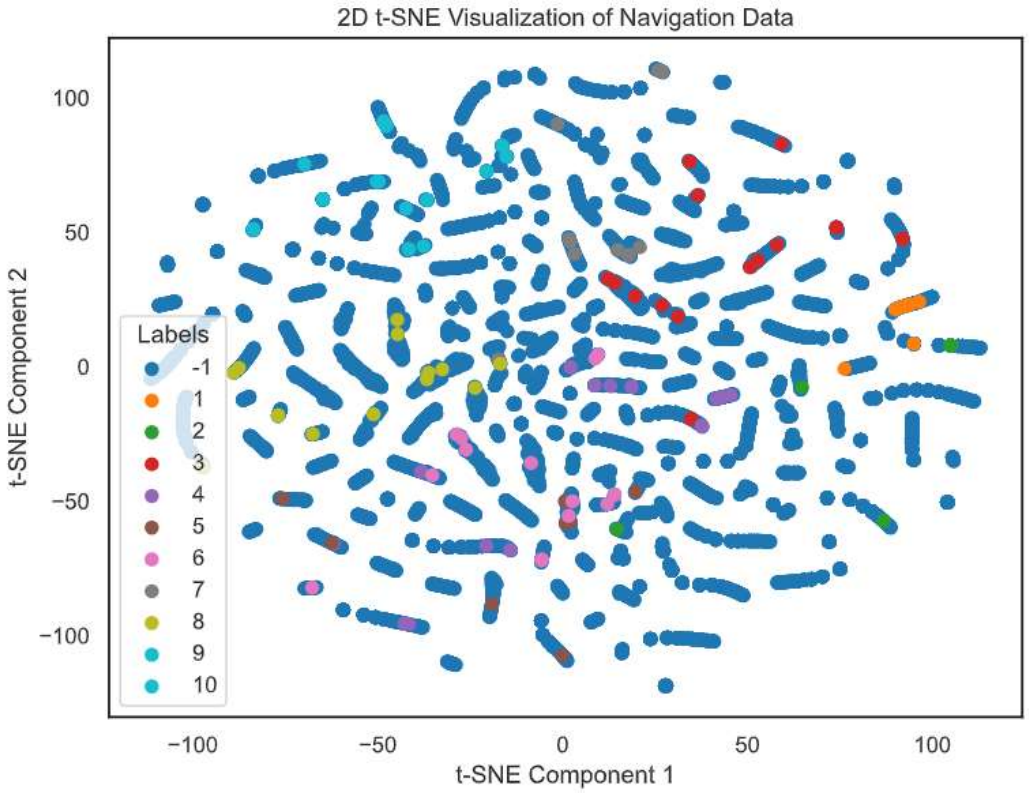
As noted, we had no labeled data on students' engagement with NAEP assessments, so it was necessary for human experts to discover such knowledge (i.e., engagement profiles). Exploration of autoencoder models with the long-short term memory layers (i.e., LSTM that preserve sequential information) led to the selection of a code space with eight dimensions. The code space, with summary statistics of total score, total response time, and the total number of item visits on the item block, a total of eleven variables, were used in clustering. Given the size of data (N=14008 by 11), we used the K-means method for easy processing. Note again that the purpose of clustering is to select representative instances for human experts' annotation and for discovery of possible engagement profiles. Other clustering methods are feasible as well.

To help human experts annotate the data, we purposely chose a number of clusters larger than necessity (in our case, the number of clusters selected was 30) to avoid missing potential engagement profiles. From each cluster, three representative instances closest to the centroid of a cluster were selected. Each representative instance is displayed as in Figure 3, as well as complimentary information about raw data sequences (such as item difficulty, item type and content), to produce a full picture of the student's engagement with the assessment for human annotation.

Human experts reviewed these representative instances, as well as extreme instances (such as highest/lowest scores, longest/shortest total times, and largest/least numbers of navigation states), aggregated and dissected the clusters and obtained ten differentiable profiles. Figure 7 is a 2-dimensional visualization of the profile distribution of about 150 initially labeled instances mapped into a 2-dimensional space, using the t-SNE techniques. Note that t-SNE is a dimensionality reduction technique commonly used for visualizing high-dimensional data in a lower-dimensional space (Van der Maaten & Hinton, 2008).

Figure 7.

Visualization of the ten profiles mapped into a 2-dimensional space



In Figure 7, the solid small dark blue circles (labeled as -1) are unlabeled instances; points with other colors are the initial labeled instances, which is about 1% of the studied sample. The order of labels (from 1 to 10) is roughly corresponding to the order of raw scores from low to high.

The preliminary ten profiles are described in Table 4. Again, refer to Figures A2 to A6 in Appendix for more examples with detailed discussion.

Table 4.

Description of the ten preliminary profiles created in the study

Label	Brief Descriptor	Profiles	Freq
1	Attempted little to no items	Unengaged group	1.77%
2	Very low score, low/regular time, and regular visit behavior	Low engagement with very low performance, navigated through most items with low time	11.02%
3	Low score, low/regular time, and regular visit behavior	Low engagement with low performance, navigated through most items with low time	14.00%
4	Low score, full/regular mixed time, and regular visit behavior	Engaged with low performance, navigated through most items, used mixed strategies	11.74%
5	Low or very low score, unregulated and/or speeded, with high visit behavior	Engaged with low performance, navigated through the items with high revisit rates, in some cases seemingly unpredictably, irregular navigation patterns with without speededness	14.16%
6	Low score, full/regular time with some prolonged item response times	Engaged with low performance, navigated through most items, spent a large amount of time on a small number of items, with or without speededness	7.67%
7	Medium score, regular time and visit behavior	Medium performing group in all dimensions	13.86%
8	Medium score, full/regular time with some prolonged item response times, and regular visit behavior	Medium performing, show strategic engagement behaviors (such as strategical response times)	18.50%
9	High score, regular time and visit behavior	High performing group, expected navigation patterns	5.43%
10	Very high score, regular time and visit behavior	Highest performing group, expected navigation patterns	1.87%

In Table 4, the very low, low, medium, high, and highest performing scores correspond roughly to the cutoffs of the lowest 10%, 1st quartile, between 2nd and 3rd quartile, 3rd quartile and the top 10% of the score range. For the last column in Table 4, please refer to the next section.

Overall, there were more profiles associated with low performing students than with high performing students. The first six profiles are low-performing ones, and they reflected different levels of engagement with the assessment from not engaged at all, low engagement, to engaged, which may reflect students' different levels of knowledge and skills, motivation in taking the assessment, affective states, time management, and/or test-taking strategies. On the other hand, the four profiles associated with medium and high scores show more engaged and expected test-taking behaviors.

Scaling Up Results – Step 3

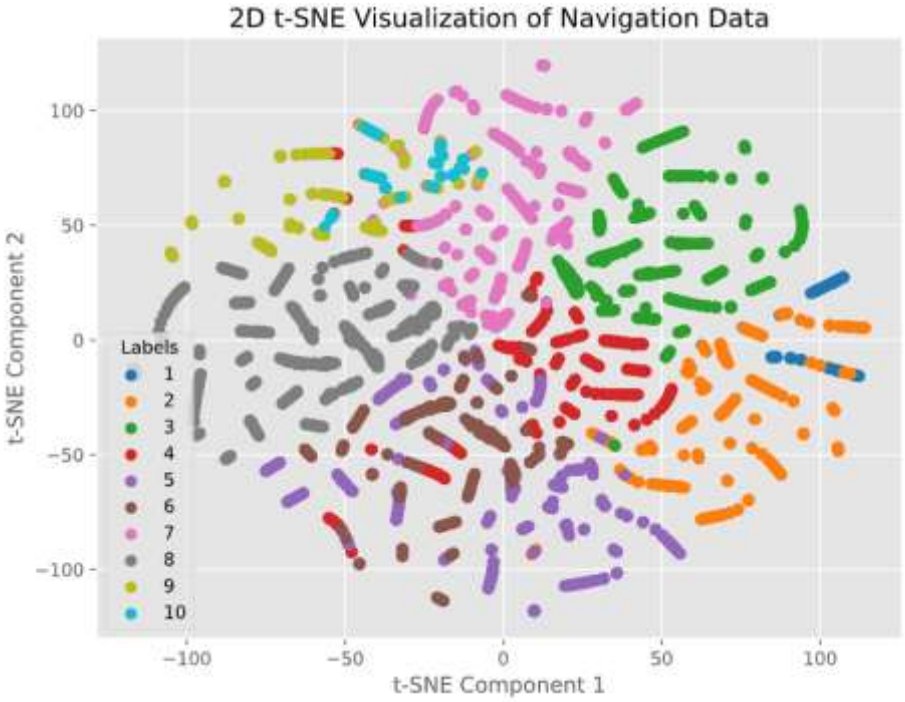
In Step-3, the ensemble model was applied to the initial labeled data in Step-2 to predict unlabeled data. Based on the model prediction, the least confident instances were selected for human manual labels, and then added to the training data. At the same time, based on the model accuracy and label confidence

trade off, instances with the pseudo label confidence larger than 0.75 were added to training data as well (users could select other thresholds to experiment). The iteration process stopped when the model accuracy could not be improved. Figure 8 shows the fully labeled data in the 2-dimensional space using the t-SNE algorithm.

The proportions of students in each preliminary engagement profile in the fully labeled data are shown in the last column of Table 4. We observed that there were very small numbers of students (about 300 out of 14008) in either Profile 1 (the unengaged group) or Profile 10 (the highest performing group), and relatively large numbers of students in the middle profiles. Overall, about 60% of students in the studied sample were in the low- or very-low-score profiles, and 40% were in the medium- or higher-score profiles.

Figure 8.

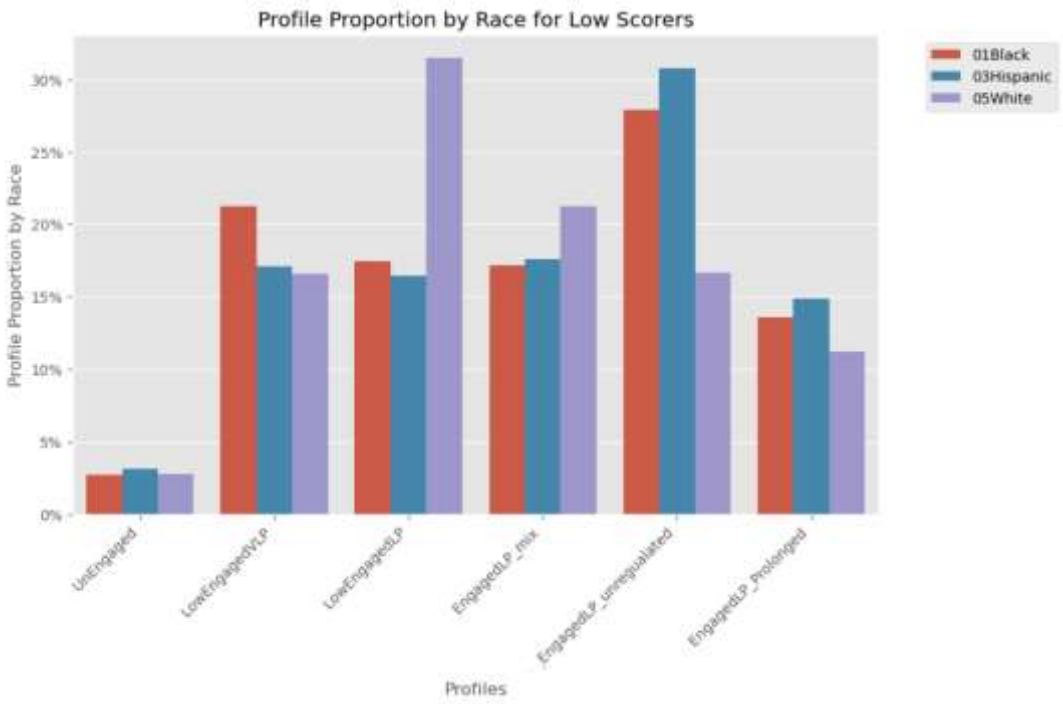
The 2-dimentional visualization of the fully labeled data, with different colors representing different profiles



These engagement profiles provide more contextualized information about test-taking processes and engagement status for individual students than the performance scores alone. Aggregation of the profiles can also shed light on student group differences. For example, among all the low performing students, Figure 9 shows that there are differences in profile proportions among different race groups (Black, n=1894; Hispanic, n=2206; and White, n=3515). A much higher proportion of white students is in Profile 3 (labeled as 'LowEngagementLP') than the black or Hispanic students, but much higher proportions of black and Hispanic students are in Profile 5 (labeled as 'EngagedLP_unregulated') than the white student group.

Figure 9.

The comparison of profile distributions by race for low performing students



Discussion

As evidenced in many recent studies (Ercikan & Pellegrino, 2017; Gordon, 2020; Guo et al., 2024; Pohl et al., 2021), digitally based assessments provide rich data about students' engagement with the assessments, which afford opportunities to investigate students' cognitive processes and problem-solving strategies, and to develop innovative assessments that better measure learning and support teaching.

In the current study, we used data from the NAEP math assessment to demonstrate how such large-scale data and AI can help generate students' engagement profiles beyond performance scores to support teaching and learning in the digital age (Office of Educational Technology, 2023). Preliminary results of the study show that there were more engagement profiles associated with low performing students, and these engagement profiles were differentiable from those with high performing students. The engagement profiles provide a holistic view of students' knowledge and skill, how they engaged with the assessment, their affective states, their test-taking strategies, and time management, etc. These profiles might suggest clues for understanding why students performed at certain levels, shed light on potential issues in their learning (such as lack of knowledge, low motivation, poor time management, difficulty with attention, focus, and organization, or other deficiency in learning and learning skills), particularly for the low-performing students (NASEM, 2018; NRC, 2000). This knowledge about students might help educators prepare differentiable intervention strategies for students in different profiles and help provide data evidence for making educational policy decision for improving teaching and learning.

Most importantly, given the large sizes of data collected from large-scale assessments, in this study, the general framework for the human-centered AI approach can support and amplify human ability in new knowledge discovery, so that useful information extraction from performance and process data can be scaled up to potentially enhance the impact of large-scale assessments. Our findings demonstrate the potential of advanced AI tools in facilitating a better understanding of students' test-taking processes and performance in context and minimizing potential false positive flags in detecting students'

engagement in existing literature (Ercikan et al., 2023; Wise, 2017). The current approach allows for the exploration of innovations in assessments through harnessing AI power in analyzing extensive educational datasets to uncover patterns, trends, and insights that help inform instructional strategies and educational policies.

The significance of our innovation in analyzing large-scale assessment data is twofold. First, the proposed generic human-centered AI architecture is applicable for mining un-labeled and partially labeled complex and large-scaled educational data for insights. Human expertise and knowledge are used in every step of the work to ensure that the results are explainable and meaningful. This architecture can help to build and accelerate the creation of large and rich benchmark data sets in education for research and practice. Second, the work takes advantage of the rich process data from large-scale assessments to explore meaningful, and potentially actionable, data-based information that may complement and enhance the impact of large-scale assessments. Students' engagement profiles with the visualizations, combined with other complementary information about the students, for example, would help educators to prepare meaningful conversations with students who have different profiles for further interventions. Aggregation of engagement profiles for groups of students within a region, a school district, or a school, would also help stakeholders to make informed educational policy decision, when compared with student bodies of similar racial/ethnic composition (NAGB, 2024b).

The current exploration work has a few limitations. First, the preliminary profiles need more refinement and improvement by involving educators and stakeholders. The second limitation is that only one item block was used from the NAEP Grade 8 Math Assessment. Further work should explore the HAI framework that can create engagement profiles across multiple item blocks and overcome the challenge of feature differences in different item blocks. Further investigation also needs to explore alternative and explainable approaches (such as new process features and different machine learning algorithms) to better capture how human experts reason to create the engagement profiles.

Declarations

Gen-AI Use: The authors of this article declare (Declaration Form #: 2611241642) that Gen-AI tools have NOT been used in any capacity for content creation in this work.

Author Contribution: The first author led the study and contributed to conceptualization, methodology, data modeling, analysis, and visualization, interpretation, and writing. All the other authors played critical roles in shaping the study by contributing to concept, methodology, data annotation, interpretation, or revision.

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

Funding: This project has been funded at least in part with Federal funds from the National Center for Education Statistics in the U.S. Department of Education. The content of the publication does not necessarily reflect the views or policies of the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endowment of the U.S. Government.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Google. https://www.tensorflow.org/

Baker, R. (2021). *Artificial intelligence in education: Bringing it all together*. In S. Vincent Lancrin (Ed.), Pushing the frontiers with AI, blockchain, and robots (pp. 43–54). OECD.

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

Ercikan, K., Guo, H., & Por, H.-H. (2023). *Uses of process data in advancing the practice and science of technology-rich assessments.* Innovating Assessments to measure and support complex skills (N. Foster & M. Piacentini, Eds.). OECD Publishing. Retrieved from https://www.oecd-ilibrary.org/education/innovating-assessments-to-measure-and-support-complex-skills_7b3123f1-en

Ercikan, K., & Pellegrino, J. (2017). *Validation of score meaning in the next generation of assessments: The use of response processes*. Routledge.

Geron, A. (2017). *Hands-on machine learning with scikit-learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media.

Gordon, E. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice, 39(3),* 72–78. Retrieved from https://doi.org/ 10.1111/emip.12370

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior, 61*, 36-46.

Guo, H., & Ercikan, K. (2021a). Differential rapid responding across language and cultural groups. *Educational Research and Evaluation, 26(5-6),* 302-327. https://doi.org/10.1080/13803611.2021.1963941

Guo, H., & Ercikan, K. (2021b). Using response-time data to compare the testing behaviors of English language learners (ells) to other test-takers (non-ells) on a mathematics assessment. *ETS Research Report, 2021(1),* 1-15. https://doi.org/10.1002/ets2.12340

Guo, H., Johnson, M., Ercikan, K., Saldivia, L. & Worthington, M. (2024). Large-scale assessments for learning: A huma-centered AI approach to contextualize test performance. *Journal of Learning Analytics, 11(2), 229-245.* https://doi.org/10.18608/jla.2024.8007

Guo, H., Rios, J., Haberman, S., Liu, O., Wang, J. & Paek, I. (2017). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education. 29(3). 173 – 183. http://doi.org/10.1080/08957347.2016.1171766*

Johnson, M. S., & Liu, X. (2022). *Psychometric considerations for the joint modeling of response and process data [Paper presentation]*. International Meeting of Psychometric Society, Bologna, Italy.

Levy, R. (2020). Implications of considering response process data for greater and lesser psychometrics. *Educational Assessment, 25(3),* 218–235.

https://doi.org/10.1080/10627197.2020.1804352

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.

Miao, F., Holmes, W., Huang, R., & Zhang, H. (2021). *AI and education: Guidance for policymakers*. UNESCO.

National Assessment Governing Board. (NAGB, 2020). *Response process data from the 2017 NAEP grade 8 mathematics assessment*. https://www.nationsreportcard.gov/process_data/

National Assessment Governing Board (NAGB, 2024a). *Mathematics assessment framework for the 2022 and 2024 National Assessment of Educational progress.* Retrieved from https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/mathematics/2022-24-nagb-math-framework-508.pdf

National Assessment Governing Board (NAGB, 2024b). *How states use and value the Nation's Report Card.* Retrieved from https://www.nagb.gov/about-us/state-and-tuda-case-studies.html

National Center for Education Statistics. (NCES, 2022). *NAEP questions tool*. Retrieved from https://nces.ed.gov/NationsReportCard/nqt/

National Research Council. (NRC, 2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. Washington, DC: The National Academies Press. Retrieved from https://doi.org/10.17226/9853

National Academies of Sciences, Engineering, and Medicine (NASEM, 2018). *How people learn II: Learners, contexts, and Cultures*. Washington, DC: The National Academies Press.

Office of Educational Technology. (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations (Report)*. Washington, DC, 2023: U.S. Department of Education.

Pellegrino, J. W. (2020). Important considerations for assessment to function in the service of education. *Educational Measurement: Issues and Practice, 39(3),* 81- 85. Retrieved from https://doi.org/10.1111/emip.12372

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. Science, 372(6540), 338-340. Retrieved from https://doi.org/10.1126/science.abd3300

Radwan, A. M. (2019). *Human active learning*. In S. M. Brito (Ed.), Active learning (chap. 2). Rijeka: IntechOpen. Retrieved from https://doi.org/10.5772/ intechopen.81371

Rios, J., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? an analysis of differential noneffortful responding on an international college-level assessment of critical thinking ISLA. *Applied Measurement in Education, 33(4),* 263–279. https://doi.org/10.1080/08957347.2020.1789141

Rios, J., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing, 17(1),* 74–104. https://doi.org/10.1080/08957347.2020.1789141

Rizve, M. N., Duarte, K., Rawat, Y. S., & Shah, M. (2021). *In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning*. In International conference on learning representations. Retrieved from https://iclr.cc/media/iclr-2021/Slides/3255.pdf

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

Wise, S. (2021). Six insights regarding test-taking disengagement. Educational Research and Evaluation, 26(5-6), 328–338. https://doi.org/10.1080/13803611.2021.1963942

Wise, S. & 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

Xie, Q., Dai, Z., Hovy, E. H., Luong, M., & Le, Q. V. (2019). Unsupervised data augmentation for consistency training. CoRR, abs/1904.12848. Retrieved from http://arxiv.org/abs/1904.12848

Zhu, X., Lafferty, J., & Ghahramani, Z. (2003). Combining active learning and semisupervised learning using gaussian fields and harmonic functions. In ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining (pp. 58–65).

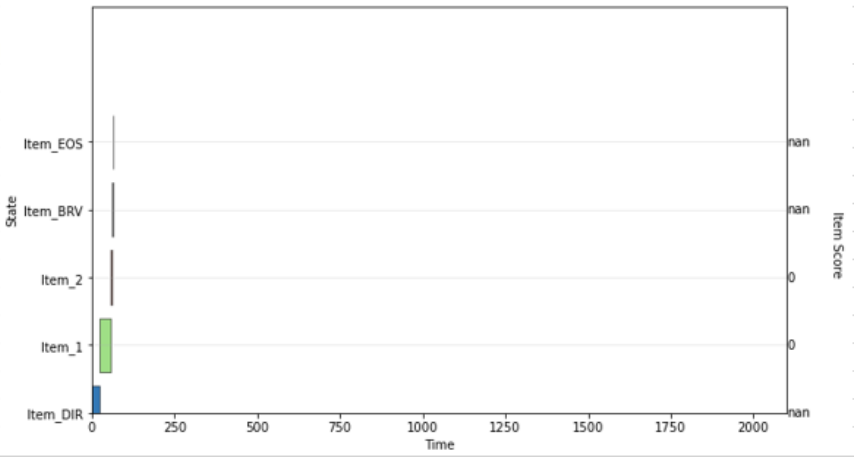
Zoanetti, N., & Griffin, P. (2017). *Log-file data as indicators for problem-solving processes*. In B. Csapo & J. Funke (Eds.), The nature of problem solving: Using research to inspire 21st century learning (chap. 11). Paris: OECD Publishing. Retrieved from https://doi.org/10.1787/9789264273955-en

Appendix

Figure A1.

One instance in Profile 1 (unengaged)

Item	Raw Score	Raw Time	Time Category	Raw Visit
1	0	36	2	1
2	0	2	1	1
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0
11	0	0	0	0
12	0	0	0	0
13	0	0	0	0
14	0	0	0	0

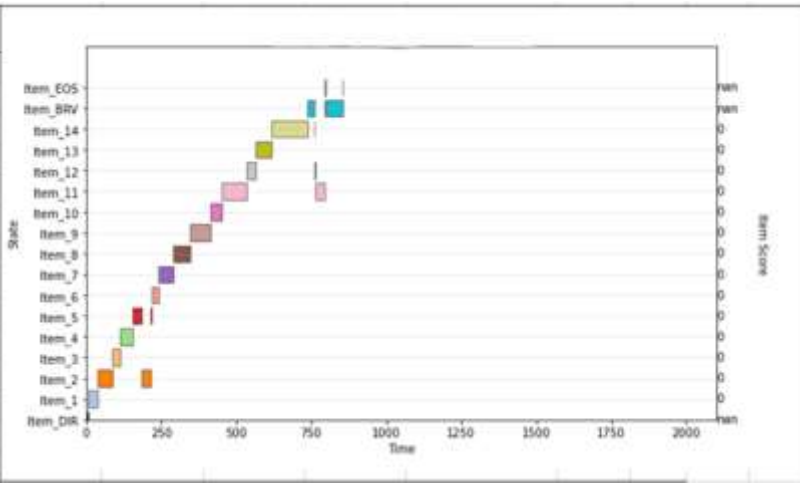


Notes. *The student had a total score of 0, a total time of 38 seconds and a total number of visit states of 5. The item summary information (item score, item response time, response time category, and number of item visits) is presented in the table on the left-hand side. The student's navigation pattern is presented in the plot on the right-hand side. This student did not engage with the assessment.*

Figure A2.

One instance in Profile 2 (Low engagement with very low score)

Item	Time	Score	Time Category	Visit
1	31	0	2	1
2	78	0	2	2
3	28	0	1	1
4	43	0	2	1
5	33	0	2	2
6	24	0	1	1
7	48	0	2	1
8	57	0	3	1
9	67	0	2	1
10	37	0	2	1
11	113	0	4	2
12	34	0	3	2
13	64	0	2	1
14	123	0	4	2

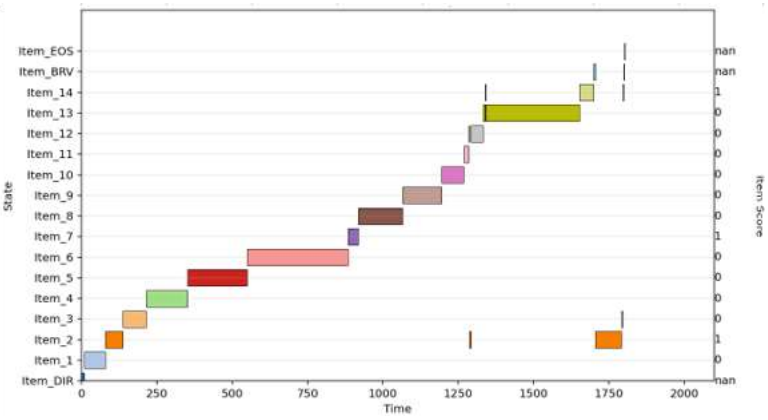


Notes. *The student had a total score of 0, a total time of 780 seconds and a total number of visit states of 25. The item summary information (item score, item response time, response time category, and number of item visits) is presented in the table on the left-hand side. The student's navigation pattern is presented in the plot on the right-hand side. This student did not engage with the assessment. The student worked through all the items but mostly without adequate effort.*

Figure A3.

One instance in Profile 4 (Low score, full/regular mixed time, and regular visit behavior)

Item	Raw Score	Raw Time	Time Category	Raw Visit
1	0	70	4	1
2	1	151	3	3
3	0	82	3	2
4	0	137	4	1
5	0	198	4	1
6	0	335	4	1
7	1	34	2	1
8	0	147	5	1
9	0	131	2	1
10	0	73	3	1
11	0	17	1	1
12	0	44	3	2
13	0	320	4	2
14	1	48	2	3

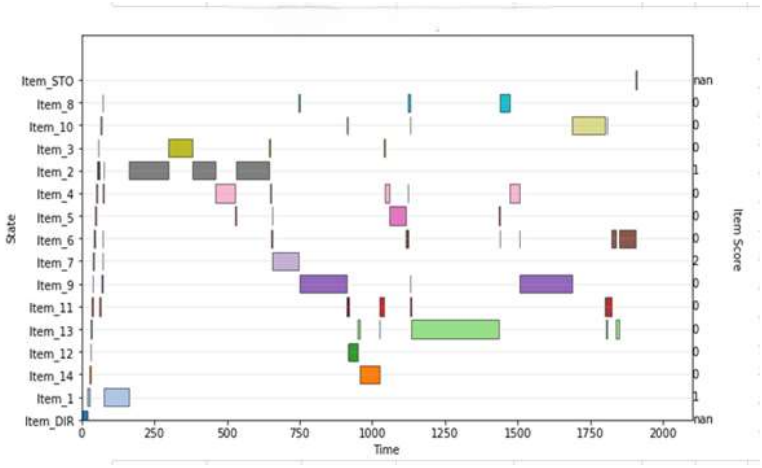


Notes. *The student had a total score of 3 out of 21, a total time of 1790 seconds and a total number of visit states of 25. The item summary information (item score, item response time, response time category, and number of item visits) is presented in the table on the left-hand side. The student's navigation pattern is presented in the plot on the right-hand side. This student used nearly full time on the item block and adopted a mixed responding strategy (relatively prolonged time on Item 8 and relatively rapid responding on Item 11, for example).*

Figure A4.

One instance in Profile 5 (Low score, unregulated and/or speeded, with high visit behaviors)

Item #	Score	Time	Time Category	Visit
1	1	96	4	2
2	1	337	5	6
3	0	89	3	4
4	0	124	4	7
5	0	70	3	5
6	0	93	2	10
7	2	94	3	3
8	0	45	2	4
9	0	342	4	5
10	0	125	4	6
11	0	45	3	8
12	0	37	3	3
13	0	320	4	6
14	0	70	3	2

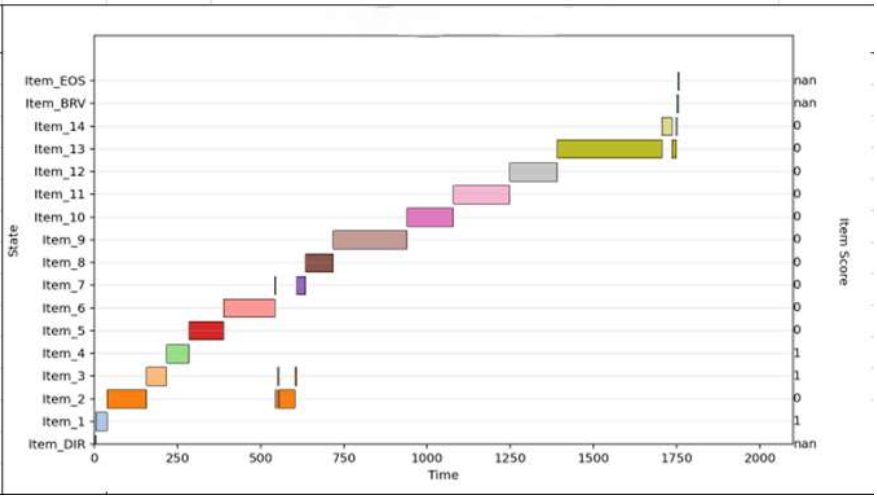


Notes. *The student had a total score of 4 out of 21, a total time of 1887 seconds and a total number of visit states of 72. The item summary information (item score, item response time, response time category, and number of item visits) is presented in the table on the left-hand side. The student's navigation pattern is presented in the plot on the right-hand side. This student visited items many times and irregularly, with a lot of item quick scanning behaviors and poor time management.*

Figure A5.

One instance in Profile 6 (Low score, full time with some prolonged item response times)

Item	Score	Time	Time Category	Visit
1	1	34	2	1
2	0	174	3	3
3	1	69	3	3
4	1	67	2	1
5	0	103	3	1
6	0	153	3	1
7	0	30	2	2
8	0	83	4	1
9	0	219	4	1
10	0	141	4	1
11	0	169	5	1
12	0	143	5	1
13	0	330	4	2
14	0	33	2	2



Notes. *The student had a total score of 3 out of 21, a total time of 1887 seconds and a total number of visit states of 72. The item summary information (item score, item response time, response time category, and number of item visits) is presented in the table on the left-hand side. The student's navigation pattern is presented in the plot on the right-hand side. This student visited items almost linearly with adequate or prolonged time effort on most items, but low performing.*