JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 11(2), 229–245. https://doi.org/10.18608/jla.2024.8007

Large-Scale Assessments for Learning: A Human-Centred AI Approach to Contextualizing Test Performance

Hongwen Guo[1], Matthew S. Johnson[2], Kadriye Ercikan[3], Luis Saldivia[4], Michelle Worthington[5]

Abstract

Large-scale assessments play a key role in education: educators and stakeholders need to know what students know and can do, so that they can be prepared for education policies and interventions in teaching and learning. However, a score from the assessment may not be enough—educators need to know why students got low scores, how students engaged with the tasks and the assessment, and how students with different levels of skills worked through the assessment. Process data, combined with response data, reflect students' test-taking processes and can provide educators such rich information, but manually labelling the complex data is hard to scale for large-scale assessments. From scratch, we leveraged machine learning techniques (including supervised, unsupervised, and active learning) and experimented with a general human-centred AI approach to help subject matter experts efficiently and effectively make sense of big data (including students' interaction sequences with the digital assessment platform, such as response, timing, and tool use sequences) to provide process profiles, that is, a holistic view of students' entire test-taking processes on the assessment, so that performance can be viewed in context. Process profiles may help identify different sources for low performance and help generate rich feedback to educators and policy makers. The released National Assessment of Educational Progress (NAEP) Grade 8 mathematics data were used to illustrate our proposed approach.

Notes for Practice

- Process data captured from the digital platforms of large-scale assessments have been researched on individual tasks or individual process features (such as students' response times, certain digital tool uses, or interactions with certain tasks), with or without using AI algorithms. Findings from process data research generally show significant associations between test-taking behaviours and performance levels, and these associations may be different for different student groups.
- The current study proposes a general and systematic human-centred AI (HAI) framework for critical decision making, to analyze complex and large data at scale, and to obtain a holistic view of students' entire test-taking processes and performance. The resulting process profiles could provide data insights, uncover new knowledge, and contextualize performance for large-scale state, national, or international assessments. Using the National Assessment of Educational Progress (NAEP) data as an illustration, the newly created process profiles could potentially enrich score reporting and provide hints for intervention.
- As a proof of concept, the current work showcases that the HAI framework can take advantage of AI tools to help human experts make sense of multi-source data (response and process data) efficiently and effectively, providing rich feedback to educators and policy makers for teaching and learning. The proposed framework has the potential to analyze other complex educational data at scale in order to generate data insights.

Keywords

Large-scale assessment, test-taking process profiles, human-in-the-loop, machine learning, deep learning

Submitted: 05/04/2023 — **Accepted:** 14/07/2024 — **Published:** 06/08/2024

Corresponding author[1] *Email:* hguo@ets.org *Address: ETS Research Institute, 660 Rosedale Rd, Princeton, NJ 08541, USA. ORCID iD:* https://orcid.org/0000-0002-1751-0918

[2] *Email:* msjohnson@ets.org *Address: ETS Research Institute, 660 Rosedale Rd, Princeton, NJ 08541, USA. ORCID iD:* https://orcid.org/0000-0003-3157-4165

[3] *Email:* kercikan@ets.org *Address: ETS Research Institute, 660 Rosedale Rd, Princeton, NJ 08541, USA. ORCID iD:* https://orcid.org/0000-0001-8056-9165

[4] Email: lsaldivia@ets.org Address: ETS Research Institute, 660 Rosedale Rd, Princeton, NJ 08541, USA. ORCID iD: https://orcid.org/0009-0007-3482-7654
[5] Email: mworthington@ets.org Address: ETS Research Institute, 660 Rosedale Rd, Princeton, NJ 08541, USA. ORCID iD: https://orcid.org/0009-0006-0480-3769

1. Introduction

As a fundamental feedback mechanism for and of learning, large-scale educational assessments allow all stakeholders to understand what students have learned and can do and where teaching resources need to be focused (Gordon, 2020; Pellegrino, 2020). Large-scale assessments routinely produce group student reports based on a performance/summary score for each student on the entire assessment. However, a performance score alone for a student from the assessment may not be enough—educators and stakeholders need to know why some students got low scores, how they engaged with the assessment, and whether and where students might need help, so that they can be better informed and prepared to improve teaching and learning. The conventional way for educators and researchers to find out such information is to examine each student's problem-solving process (such as show your work) or use interviews of students in cognitive labs (such as think-aloud protocols). When the student group is large, this is a formidable task. This is particularly challenging when students take large-scale assessments such as those at the state, national, or international level.

With technology advances, many large-scale assessments have been administered on digital testing platforms, including online testing environments (National Assessment Governing Board, 2020; Organisation for Economic Co-operation and Development, 2020). Digital-based assessments generate a huge amount of data by automatically logging every action a student took on the digital platform. These log files, when appropriately processed and analyzed, can be used to validate score meaning and improve test design. For related studies and discussions, interested readers can refer to recent books such as Ercikan and Pellegrino (2017) and Foster and Piacentini (2023). In addition, students' profiles based on process data can potentially enhance score reporting and help to generate rich feedback for educators to communicate with their students in a more meaningful way, for policy makers to make better-informed decisions, and for assessment developers to improve assessment tasks and designs. More important, the accessibility of machine learning (ML) and deep learning algorithms and other AI tools makes it promising to understand students' test-taking and problem-solving processes at scale on large-scale assessments, which could help accelerate next-generation assessment development to measure creative thinking, problem-solving, and other higher-level skills (Ercikan et al., 2023).

Log files and process data have been analyzed in the digital learning space with or without ML algorithms and AI tools. For example, Lagud and Rodrigo (2010) investigated the relationship between students' affective profiles and their learning profiles in an intelligent tutoring system (ITS) for algebra. A student's affective profile was the distribution of affective states (such as boredom, confusion, delight, flow, and frustration) observed in field observations. The student's learning profile, on the other hand, was based on the number of correct answers, the highest task difficulty level, the average time on the tasks, and the average number of steps to solve a task, some of which were collected from log files. Using a one-way analysis of variance (ANOVA), the authors found, for example, that high-performing students' experience of affect states was different from that of low-performing students, who had minimal correct responses and spent the longest time on solving algebra problems. Students who took the longest time or used the greatest number of steps to solve the algebra problems experienced the most confusion. A persistent or prolonged state of confusion is associated with negative learning experiences that can lead to students' disengagement and frustration (R. Baker et al., 2010; Nawaz et al., 2020). Paquette and Baker (2019) compared two commonly used approaches, knowledge engineering and ML, to model students' game-the-system behaviours in an ITS. Such behaviours mean that the learners may abuse the learning environment's support functionalities to succeed by guessing or copying answers. They found that machine-learned models required fewer resources to develop but were less interpretable and generalizable than with the knowledge engineering approach. In the ITS for science learning, which uses the learning-by-teaching paradigm, Biswas and colleagues (2016) showed that using different approaches (hidden Markov models, sequence mining, and statistical analysis) on multi-modal data could differentiate successful and unsuccessful students, identify plausible sources of mistakes, and produce different profiles (such as frequent researchers and careful editors, strategic experimenters, confused guessers, etc.). These findings in turn help to improve the learning system to better assist students. Nawaz and colleagues (2020) also studied log files in a simulation-based predict-observe-explain (POE) environment. They compared different groups using the *Student T test* and found that indicators extracted from log data, such as task attempts and time on task, possibly reflected the moments of confusion. Students who had high confidence but made errors showed signs of confusion, with longer time on the task and more attempts. In an educational gaming environment, Kleinman and colleagues (2022) presented a human-in-the-loop approach to allow a stakeholder to better understand both the data and the analysis algorithms. The visualization tool in their study helped to show how students' individual problem-solving sequences differed from optimal ones. Using a learner's history of past interactions with an ITS, Gervet and colleagues (2020) compared deep learning models with classical Markov process and logistic regression models to predict students' future performance. Their results showed that deep knowledge tracing performs best when data sets are large or when precise temporal information matters most. R. Baker (2021) provided an overview on progress in learning analytics using log files from ITSs. Learning analytics research has shown, in many real-life examples, that log data collected from ITSs improve learning by contextualizing task results; shedding light on students' problem-solving processes; and providing timely feedback to students, educators, and designers (R. Baker et al., 2019; Biswas et al., 2016).

In light of the progress in the digital learning space, research on log data collected from large-scale assessments may have a significant impact at a larger scale on teaching and learning by providing data insight for educational policy changes at international, national, and state levels, and through helping educators to integrate assessment, curriculum, and instruction in education (Gordon, 2020). Researchers in the educational measurement field have in fact been using log/process data, with or without AI algorithms, to address a variety of measurement questions in assessments. For example, time-on-task, or response time, one of the most commonly used process features extracted from log data, is often used for descriptive and diagnostic purposes to provide information on test-taking strategies (such as rapid-guessing (RG) behaviours), score validity on the assessments, and measure its relationship with performance (Ercikan et al., 2020, 2023; Greiff et al., 2016; Guo & Ercikan, 2021a; Pools & Monseur, 2021; Wise, 2021). After analyzing an interactive problem-solving task, Zoanetti and Griffin (2017) discussed a theoretical rationale and an empirical methodology, which showed the potential of using process sequence data to develop rules for scoring complex problem-solving tasks. Using sequence-mining techniques, studies demonstrate how test-taking behaviour patterns were associated with incorrect responses, which may help us understand different sources of mistakes for an individual problem-solving task (Ulitzsch et al., 2022). Using information extracted from keystroke data, Markov processes and ML algorithms were used to analyze testing modes (Guo, 2022) and to model writing state and duration time sequences to help provide useful information and feedback on students' writing styles and for classroom instruction (Bennett et al., 2022; Deane et al., 2021; Guo et al., 2020). In addition, researchers have attempted to incorporate process data in psychometric models to yield inferences on assessments (Johnson & Liu, 2022; Levy, 2020).

While the above-cited efforts in assessment research have made notable contributions toward identifying and understanding process data collected from digital-based assessments, we still need to deepen our understanding of the complex and big educational data produced from large-scale assessments. The current study has four major differences from previous studies on assessment data. First, the aforementioned studies mainly provide snapshots of the assessment data by focusing on either individual process features (such as response time) or action sequences on an individual task to provide insight into students' engagement with the assessment. In the current study, we focus on students' entire test-taking processes and performance on the assessment, including both item-level (i.e., task-level) sequential data and test-level information, to create a rich process profile for each student. Second, previously, when process features were investigated, most researchers used static process features and ignored the sequential information among items. When the sequential information (such as sequences of actions) on an individual task was analyzed, the associated time duration of each action was usually ignored, which may lead to difficulties in data interpretation (please refer to Section 4.1.2 below for examples and discussions). In the current study, we consider the sequential information in item responses and item response times, as well as that in digital tool uses.

Third, when ML algorithms were used to analyze process data collected from educational assessments, supervised learning approaches were often used in previous studies. That is, process features were often used to predict item scores (correct or incorrect answers), test scores, or available statistics from the assessment. When unsupervised learning algorithms were used, the meaning of the resulting clusters was usually hard to interpret. The current study's goal is to understand students' entire test-taking processes and performance, and thus both process data and response data (also referred as multi-source data in the later discussion) are integrated and analyzed to shed light on how students performed, how they managed their time, and how they used the digital tools in the test-taking process. Such complex sequential process data have not been analyzed as an integrated data source. Hence, to make sense of such multi-source data, humans should be in the loop of the analysis and annotation processes. Moreover, annotating the huge amount of data from large-scale assessments is time-consuming and labour-intensive, let alone concerns about label quality from crowd-sourcing if it could be done. Therefore, we explored a general AI-augmented active ML approach to assist human experts, to accelerate knowledge discovery in the rich assessment data, and to assist in data interpretation. Last but not least, a difference of the current study from previous ones is the important consideration in building transferable knowledge and models from one representative sample to the other samples for the large-scale assessments. For example, in 2017, about 150,000 eighth-graders were sampled from the nation to participate in the National Assessment of Educational Progress (NAEP) Grade 8 mathematics assessment, with each student working on one of roughly 50 different test booklets. However, process data from only one representative sample were released (i.e., about 3,000 students who worked on the released test booklet; please refer to Section 3 below for more details). It is critical that process profiles created from the released sample help process profile creation for students in the other samples when more data are available/released.

Hence, our goal in the study is to harness the power of both AI and big data from large-scale assessments to develop a general framework to make sense of the complex data and create meaningful process profiles on students' entire test-taking processes. Specifically, the current study addresses the following research questions:

RQ-1. How can we prepare data so that we can not only preserve the rich information in the multi-source data but also take advantage of the survey design of large-scale assessments to obtain possibly transferable knowledge?

RQ-2. How can we help subject matter experts (i.e., human experts) discover data insights and accelerate the creation of process profiles at scale?

To assist with our exploration and help answer the above research questions, we used the released data from the NAEP for illustration. The annotated data provide different process profiles that may shed light on students' performance and engagement with the NAEP assessment and provide meaningful feedback for stakeholders to make informed decisions in education and for educators to facilitate discussions with students; the findings may help to prepare intervention if necessary.

The paper is structured as follows. In Section 2, we introduce the architecture of our proposed human-centred AI (HAI) approach of annotating complex assessment data from scratch. In this approach, we adopted the human-in-the-loop AI practice in every step: from data preprocessing to ML model training. In the next data section, we describe the 2017 NAEP mathematics assessment, its testing environment, and data collection. We also introduce both the performance data and the process data. In Section 4, we demonstrate how we used self-supervised and unsupervised learning algorithms to help human experts have a cold start on the unlabelled data, and then how we used the active learning algorithm, integrated with semi-supervised learning, to iterate through the unlabelled data, with human experts' input, to annotate the entire data set. Findings are presented in the results section. In the discussion section, we summarize the procedures in the study, speculate on the potential uses of the labelled data, and discussed the implications and limitations of the current study. We also discuss the possible future directions in making use of AI tools and assessment data for assisting teaching and learning.

2. HAI Framework

The application of AI in education has developed in multiple directions, from designing AI tools to support students' learning and assessment to designing AI support to teachers in classroom and AI systems for management of educational institutions (T. Baker et al., 2019). HAI applications have been strongly recommended in education to be based on established, modern learning principles; wisdom of educational practitioners; and expertise in the educational assessment community (Office of Educational Technology, 2023; Miao et al., 2021).

In this study, the application of HAI is to amplify and augment (rather than displace) human expertise in understanding students "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).

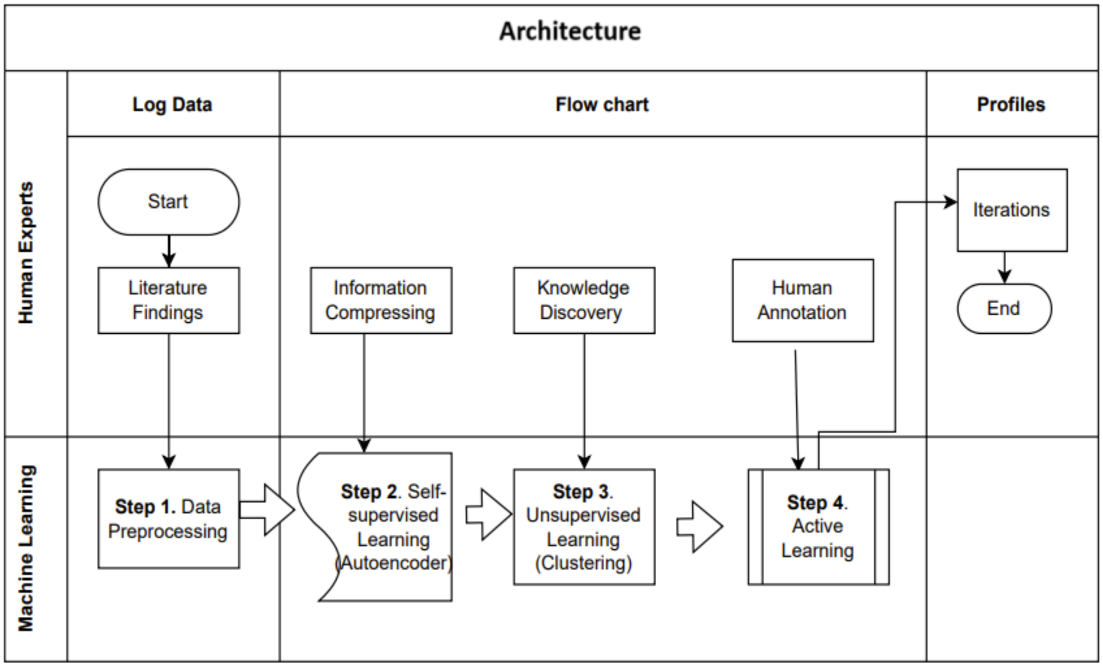


Figure 1. The HAI profiling architecture.

The proposed architecture (Figure 1) of the HAI approach attempts to maximize AI power and minimize redundancy of human labour, so that the data can be effectively and efficiently annotated to address big-data challenges.

There are four major steps in the architecture in Figure 1. The *first step* is preprocessing data. This step includes data cleaning and feature generation guided by knowledge obtained from previous research on process data and test-taking behaviours from large-scale educational assessments, to make sure that the input data to ML models are easily understandable, interpretable, and meaningful. In addition, feature comparability across different test takers and testing occasions is considered to address RQ-1 (please also refer to Section 4 on data preprocessing). The *second step* is to use an autoencoder (a self-supervised deep learning model) to compress the input sequential data (item responses, response times, and tool uses on the entire test) into a low-dimensional space (also called the code space). Autoencoders are artificial neural networks capable of learning dense representation of the input data, and they are self-supervised (or unsupervised) because input data are not labelled (Geron, 2017, Chapter 17). A successful autoencoder can recover the input data after decoding the code space. Layers that preserve sequential information are used in the autoencoder to address both RQ-1 and RQ-2 by preserving sequential information for clustering (please also refer to Section 4.2). Step 2 helps us create a common code space which could be used when we analyze other samples on different booklets in the future if data are available, because different booklets had different items and test lengths.

The *third step* in the architecture uses clustering, an unsupervised ML method, on the code space, in addition to test-level information, to find clusters and a few representative instances in each cluster. In this step, a large number of clusters is chosen on purpose to help with knowledge discovery. The subject matter experts examine the representative instances in each cluster, assign labels, and decide whether some clusters can be aggregated into one process profile based on their examination, assisted by data visualizations and their content expertise. By visualizing and analyzing these clusters, human experts can gain insight into the preliminary profiles that have emerged from the data. Human experts play a crucial role in this step, as they create and provide annotations to the resulting profiles and add their domain knowledge to the interpretation process. This step helps to address knowledge discovery in RQ-2.

In the *fourth step*, the active learning and semi-supervised learning (SSL) approaches are integrated. There are two components in active learning: an ML model and a group of human experts. The ML model is initiated with the small number of labelled data obtained in the third step, and then it queries the unlabelled data and selects (usually a small number of) informative and challenging instances for the human experts' annotation and feedback. The newly labelled data are then used to rebuild the ML model. The active learning approach iterates the query and learning processes until a criterion is met. The iteration also ensures that the model focuses on the most informative and critical samples and obtains human experts' inputs for re-training and upgrading the model, further improving its performance. Please refer to the flowchart in Figure 2 (Radwan, 2019). Step 4 in the architecture helps address scaling up in RQ-2 to empower human experts to build the AI models. Note that human experts' involvement in the first two steps is more related to decision making in data preparation, and their involvement in the last two steps is directly related to profile creation.

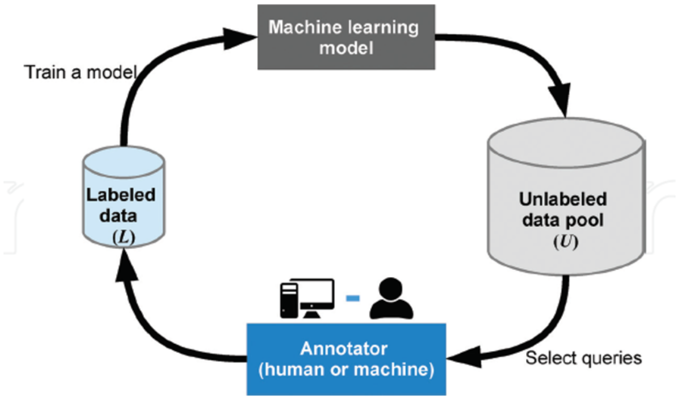


Figure 2. The iterative active learning process integrated with semi-supervised learning.

Procedures for each of the four steps are task-specific, and details will be presented in the data annotation process section that follows. This HAI framework ensures that the human experts are actively involved throughout the analysis process, allowing them to interpret data and benefit from AI assistance.

3. Data

For illustration purposes, we used the released two mathematics item blocks in the 2017 NAEP Grade 8 mathematics assessment, which measured students' knowledge and skills in mathematics and students' ability to apply their knowledge in problem-solving situations. The NAEP mathematics assessment covered five broad content areas: number properties and operations; measurement; geometry; data analysis, statistics, and probability; and algebra. The items on the assessment had three levels (low, moderate, and high) of complexity based on demands on students' thinking, and they were presented in four item types: multiple choice (MC), composite selected response (SR), short constructed response (SCR), and extended constructed response (ECR).

In 2017, the NAEP mathematics assessment was administered for the first time as a digital-based assessment (DBA; National Assessment Governing Board, 2020) in Grades 4 and 8. This allowed for the collection of process data, including information on how long students spent on the assessment questions and how they used onscreen tools to develop their responses. At the time of conducting this research, the 2017 process data are the only publicly available process data released from the NAEP assessments.

About 145,000 Grade 8 students participated in the 2017 NAEP mathematics assessment. Data from one test form (booklet with two item blocks), with about 3,000 students, were released and available for secondary analysis (National Assessment Governing Board, 2020). These data contain both students' performance data (i.e., item responses) and process data that reflect students' interaction with the NAEP testing environment.

3.1 Performance Data

Table 1 shows the item information on the two released item blocks. The first item block (8M9) in the form has 19 items[1], and the second item block (8M3) has 15 items. Students have 30 minutes to work on each item block.

Table 1. The item information of the released blocks.

Seq	Block	Type	Difficulty	Max Score	Short Content description
1	8M9	MC	Medium	1	Measurement
2	8M9	MC	Easy	1	Geometry
3	8M9	MC	Medium	1	Numbers
4	8M9	SR	Easy	1	Numbers
5	8M9	MC	Hard	1	Data and Stats
6	8M9	MC	Hard	1	Geometry
7	8M9	SCR	Easy	2	Algebra
8	8M9	MC	Hard	1	Measurement
9	8M9	SCR	Hard	1	Geometry
10	8M9	MC	Hard	1	Numbers
11	8M9	MC	Medium	1	Algebra
12	8M9	SCR	Hard	2	Numbers
13	8M9	SCR	Hard	1	Algebra
14	8M9	MC	Hard	1	Data and Stats
15	8M9	MC	Medium	1	Algebra
16	8M9	MC	Hard	1	Numbers
17	8M9	MC	Medium	1	Algebra
18	8M9	MC	Hard	1	Data and Stats
19	8M9	MC	Easy	1	Algebra
1	8M3	MC	Easy	1	Numbers
2	8M3	SR	Easy	1	Data and Stats
3	8M3	SCR	Medium	1	Numbers
4	8M3	SR	Easy	2	Algebra
5	8M3	SR	Easy	2	Measurement
6	8M3	SR	Medium	1	Algebra
7	8M3	SCR	Hard	1	Measurement
8	8M3	MC	Hard	1	Geometry
9	8M3	SR	Easy	1	Measurement
10	8M3	SR	Hard	2	Geometry
11	8M3	SCR	Hard	2	Algebra
12	8M3	SR	Hard	2	Geometry
13	8M3	ECR	Hard	4	Data and Stats
14	8M3	SR	Hard	2	Data and Stats
15	8M3	SCR	Hard	2	Geometry

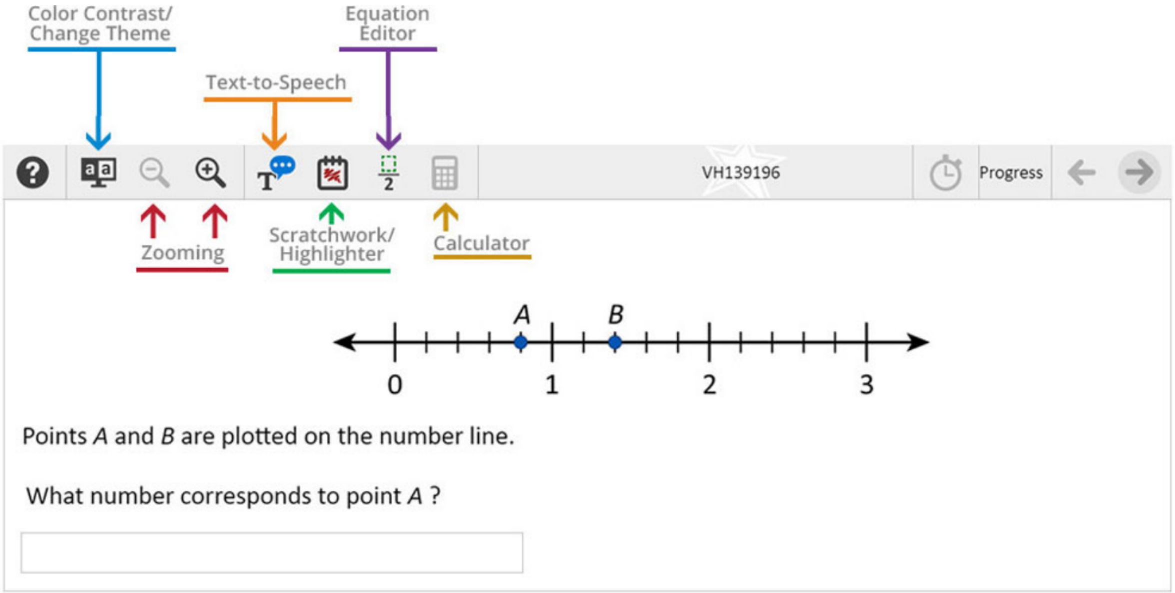
[1] Note that the content for the last four items in 8M9 was not released.

The items have varying item difficulty and are presented in four item types (MC, SR, SCR, and ECR) to cover the five content areas. The maximum item score varies from 1 point to 4 points, and the total maximum raw score on the test form is 45. For example, item 13 in block 8M3 (refer to Figure 8 in the Appendix) is an ECR 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 all correct responses. Detailed information on the content of the release items and scoring rules can be found on the NCES website (National Center for Education Statistics, 2022).

3.2 Process Data

NAEP digital-based assessments offer a testing environment that makes it possible to record students' interactions with the digital platform when they solve the tasks. Figure 3 shows a screenshot of the platform and digital tools that a student could use (National Assessment Governing Board, 2020).

Starting from the upper left of the screen in Figure 3, the digital tools include help (a question mark, which opens up a page to describe the use of available tools), colour contrast and change theme, zoom in/out, text-to-speech, scratchwork/highlight (which opens up a set of scratch tools for drawing, highlighting, and erasing marks), equation editor (for onscreen math keyboard), and calculator. On the upper right of the screen, students can monitor their session time (a clock icon, which shows how much time is left in the session), check their progress (which shows what items they have and have not responded to), and navigate between page and items (the forward and backward arrows). Please refer to the NAEP math tutorial for details (National Center for Education Statistics, 2021). Students' interactions with the testing environment were logged and collected to produce process data.



Points *A* and *B* are plotted on the number line.

What number corresponds to point *A* ?

Figure 3. The NAEP testing environment.

The released process data contain logs of the response process data collected from each student, such as item response time, number of times that a student opened the onscreen calculator, number of times that a student used the highlighter tool, etc. Each student was associated with 16 process features on each math item, in addition to other background variables. Therefore, on the released booklet, each student has 544 data points related to item score, item response time, and different digital tool uses. Because of space limitations, please refer to National Assessment Governing Board (2020) for detailed information on the process data variables.

4. Data Annotation Process and Methods

Using the released data on the two mathematics blocks (a booklet) as an illustration, we started the annotation process from scratch, because simultaneously annotating students' entire test-taking processes and performance (i.e., item responses, item response times, and digital tool uses) has not been attempted before on large-scale assessments. In the following subsections, we provide detailed information on each of the four steps in the architecture in Figure 1.

4.1 Data Preprocessing and Feature Engineering

4.1.1 Features

In preparing the data for analysis, three sequences of features were used. The first feature sequence is the item score sequence, where 1 indicates a correct answer and 0 a wrong or no answer on a dichotomous item. A few items were polytomously scored, from 0 to 2 or 4; one such item is shown in Figure 8.

The second feature sequence is the item response time sequence. As shown in the literature (Ercikan et al., 2020; Guo & Ercikan, 2021a; Rios et al., 2017), item response times have been shown to be significantly associated with performance on large-scale assessments. To solve an item, an appropriate amount of time needs to be spent on understanding the question and finding a solution. A rapid response, on the other hand, indicates students' lack of meaningful engagement with an assessment situation, such that they respond to assessment items rapidly without taking the time to read and fully consider them. A rapid response is likely a random guess, which adds noise to response data and does not reflect students' true knowledge and skills (Wise, 2017, 2021; Ercikan et al., 2023). To identify the RG response, many methods have been developed that use a theory-guided approach, a data-driven approach, or a mixture of the two. In this study, we used the hybrid procedure to flag the rapid responses proposed in Guo and Ercikan (2021b), an extension of that used in Rios and Guo (2020), which selects the minimum time threshold produced from three methods for an individual item, so that the results are conservative and reduce Type I error. In addition, because a small difference (such as one second or two seconds) between students' response times may not be meaningful, we discretized timing data in this step, where $0,1,2,3,4$, and 5 stand for no time, RG, between RG and median, between median and the third quartile, between the third quartile and the 95th percentile, and above the 95th percentile (prolonged duration), respectively. This discretization helped human experts understand a student's relative standing in the studied group that took the released blocks in terms of item response times, and it also helped to produce comparable variables across different item blocks when annotating all students who took the NAEP assessment. This discretization makes the time feature comparable across different student groups that took different test forms because of the group equivalency by design (National Center for Education Statistics, 2017). We asked human experts to pay extra attention to response time categories $T = 0,1$, and 5, where 0 indicated a not-reached item, 1 rapid responding, and 5 an extremely long time spent on the item, since excessive response time may be associated with difficulty in solving a question (Greiff et al., 2016). For example, item 13 in block 8M3 (refer to Figure 8 in the Appendix) is a hard item, and its median response time is 168 seconds. Any non-zero response time less than 97 seconds is labelled as 1 (rapid response), and any response time longer than 356 seconds is labelled as 5.

The third feature sequence is the tool use sequence. The tool use variable was an aggregation across all available digital tool use counts on an item for each student because it was observed that the tools were rarely used by most of the eighth-grade students. ToolUse was the discretized version of the tool use variable, where $0,1,2$, and 3 indicate no use, use in the lower quartile, use between the first and 95th percentile, and use above the 95th percentile, for reasons like those in discretizing response times. For example, on item 13 in block 8M3 (refer to Figure 8), the 95th percentile of the tool use is 3, and any value of tool use larger than 3 is labelled as 3.

4.1.2 Input Data Formatting

In data preparation, we emphasized preserving sequential information, because one isolated event is open to multiple interpretations as to what generated it. For example, RG behaviours observed at the beginning of the assessment and those observed at the very end of the assessment under time pressure clearly contain different meanings: the earlier RG behaviours are likely to indicate low test-taking motivation, and the latter ones are likely to indicate test speededness. Similar arguments apply to digital tool uses: a large number of tool uses at the beginning of the test session may indicate curiosity and exploration, while a large number of tool uses at the end of the test session may indicate tool playing to pass the time. Table 2 shows the input matrix/tensor of one student for step 2. Note that both time and tool sequences are discretized.

4.2 Information Compression

When flattening the input matrix, principal component analysis (PCA)[2] showed that there were three or four principal factors, so in the autoencoder model, the latent/code space dimension was chosen to be three (experiments with four dimensions did not show significant differences). Autoencoder is a type of self-supervised learning model that can learn a compressed representation of input data to capture the essential latent factors (Geron, 2017). To preserve the sequential information in the input data (i.e., the order of when items were presented), we used long-short-term memory (LSTM) layers, a special case of the recursive neural network (RNN), in the encoding and decoding layers. Please refer to Figure 4, modified from Trinh and colleagues (2019). Consideration of the sequential information may help to differentiate actions and events in context, as discussed earlier. One isolated action/event is open to multiple interpretations depending on where it occurred and how much

[2] PCA is a statistical technique used for dimensionality reduction. It aims to transform a data set consisting of possibly correlated variables into a new set of uncorrelated variables called principal components to preserve the most important information.

Table 2. The input data of one student in step 2.

Score	Time	Tool
1	4	0
1	2	0
1	2	1
1	2	1
1	2	1
0	4	0
2	3	0
0	5	1
1	2	0
1	2	0
1	3	0
2	3	0
1	3	2
1	2	2
1	2	0
1	4	0
1	5	0
1	3	0
1	2	0
1	2	0
1	2	0
0	3	0
2	2	0
1	2	0
1	3	0
2	4	0
2	4	0
2	4	0
4	5	0
2	5	0
1	2	0

time was left (the early items, the items in the middle of the session, or the latter items presented in a session). In addition, information compression in this step also helps to fix the dimension of the code space for building a general ML model when we need to analyze other item blocks because the test length varies for different item blocks.

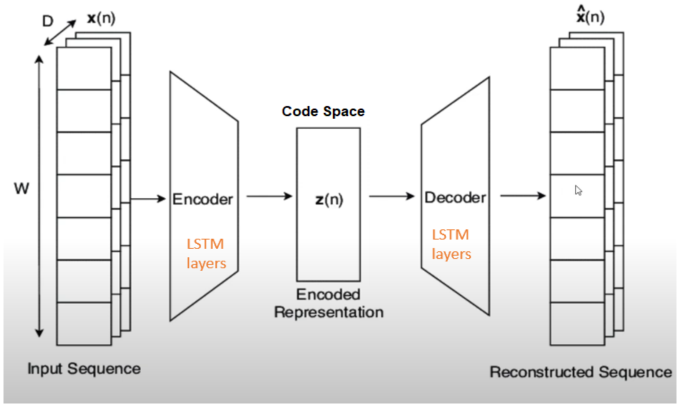


Figure 4. The LSTM-based autoencoder presentation in step 2.

4.3 Knowledge Discovery

Because we had 0 labels on the complex data, it is a necessary step to include knowledge discovery to guide ML models. For this, we applied cluster analysis on the latent space produced in the autoencoder step. The Gaussian mixture method was chosen to avoid the tendency of equal-sized clusters from other commonly used clustering methods such as K-means.

A larger number of clusters than necessary was chosen on purpose (in our case, $C = 30$) so that human experts could look into the representative sequences from each cluster, understand and interpret each cluster, and split or aggregate clusters when necessary. To help human experts annotate the clusters, the three representative sequences closest to the centroid were selected from each cluster. The representative sequence was also embedded with raw data sequences, information on item block (total block time, total score, and total tool use), and item statistics (such as item difficulty, item type, and content type) to produce a full picture for their annotation of the sequences. When human experts evaluated the data of a representative student, they not only looked at the data on each item but also evaluated where it occurred, as well as the student's overall score, time, and tool uses on the test. The information provided to the human experts is displayed in the examples in the Appendix from Table A.1 to Table A.11. For example, a rapid response occurring on the first few items in a session may indicate low motivation (Table A.1; item 2 in block 8M9). When it is in the middle of the session, it may indicate a test-taking strategy of skipping a difficult question (Table A.5; item 12 in block 8M9). On the other hand, a rapid response occurring on the last few items in the session may indicate that the student was running out of time and under time pressure (i.e., the test was speeded to them), particularly when the total session time was expiring (Tables A.6 and A.7 in block 8M9).

4.4 Active Learning Integrated with Semi-Supervised Learning

In the active learning step, our approach was integrated with semi-supervised learning; this integration has been shown to achieve state-of-the-art results (Xie et al., 2020; Zhu et al., 2003). The ML model we used in this step was an ensemble of random forest classifiers and support vector machine (SVM) classifiers. The random forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree, whereas SVM performs multiclass classification with multiple binary classifiers and selects the class that won the most duels (Geron, 2017).

A soft voting classifier based on the ensemble (refer to Figure 5) produced a prediction of an instance's class. Pseudo-labels with high confidence, which also maintained high model accuracy, were added to the training data, together with the human experts' newly labelled data from the query, to rebuild the ensemble model. This process iterated until a criterion was met (please refer to Figure 2). In our case, the criterion was that all instances in the sample were labelled by human experts or machine.

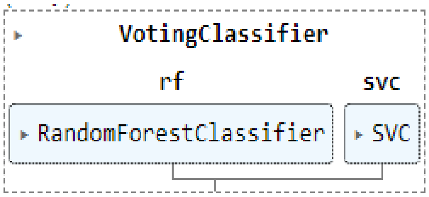


Figure 5. The ensemble classifier used in step 4.

5. Results

The sample used in the illustration contained about 2,500 students (after cleaning out irregular process data), which were representative, by design, of the student population that participated in the 2017 NAEP Grade 8 mathematics assessment.

Input matrices/tensors of students' sequential data were prepared to feed into ML methods. We used *TensorFlow* (Abadi et al., 2015) and its associated libraries to implement all of the ML methods used in the analysis.

5.1 Student Process Profiles Based on Sequential Data

Based on the clustering analysis in step 3 (knowledge discovery), human experts manually annotated 90 representative students' sequential data, three from each of the 30 clusters. Because of space limitations, examples of the visualization of representative instances to assist human experts' annotation are provided in the Appendix. Clusters were then aggregated and dissected to obtain 11 meaningful profiles, agreed upon by our human experts. Figure 6 is a visualization of the profile distribution mapped into a two-dimensional space, using the first two principal components of the code space. Note that in Figure 6, the solid small circles (labelled as -1) are unlabelled instances; points with other symbols and colours are the 90 labelled instances, which is about 3% of the studied sample. The order of labels (from 0 to 10) roughly corresponds to the order of raw scores from low to high. We noticed that in the two-dimensional plot, the profiles associated with very low and very high scores are more distinctive than other profiles.

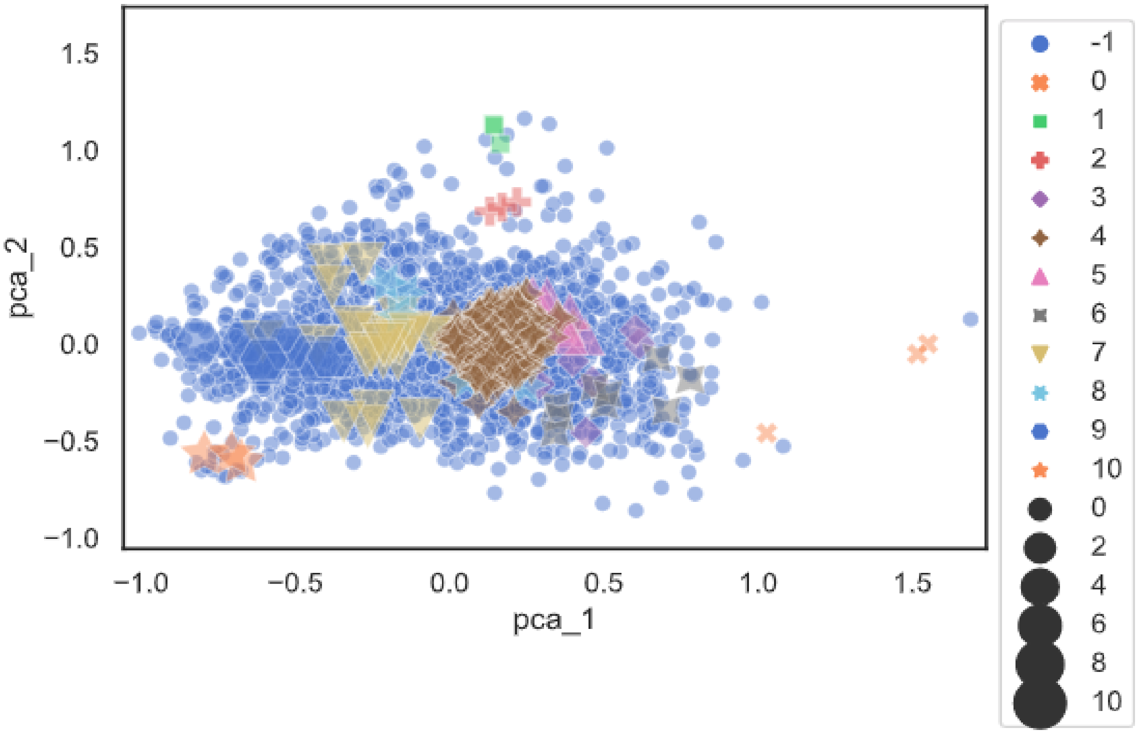


Figure 6. Visualization of the 11 profiles mapped into a two-dimensional space in step 3.

Table 3 shows the preliminary process profile descriptors and labels. Examples in the Appendix also include detailed discussion of some of these process profiles. In the table, very low, low, medium, high, and highest performing correspond roughly to the lowest 5%, below second quartile, between second and third quartile, between third quartile and the top 5%, and the top 5% of the score range. For the last column in the table, please refer to Figure 7 and its discussion.

Table 3. The resulting 11 preliminary profiles in step 3.

Label	Process Profile	Rel Freq
0	Unengaged (attempted little to no items)	0.27%
1	Low performing and disengaged, with some tool play (low time on the second block, and high tool use)	0.92%
2	Low performing, with regulated time but potentially distracted by tools (regulated time, and a lot of tool uses)	2.08%
3	Very low or low performing but overall engaged, with regulated time and regulated tool uses (some prolonged item times)	7.31%
4	Low performing and disengaged, with regulated tool uses (low time on the second block, may use mixed responding strategies)	29.54%
5	Low performing, potentially speeded, and struggling, with moderate to high tool uses (speeded on the first block)	7.19%
6	Low performing, potentially speeded, and struggling, with low tool uses (possibly speeded on both blocks)	8.38%
7	Medium performing in all dimensions (score, time, tool use)	27.88%
8	Medium performing and potentially speeded, with high tool uses (speeded on the first block)	3.15%
9	High performing and regulated time and tool uses	12.12%
10	Highest-performing group and nearly full time, with low tool uses	1.15%

5.2 Active Learning with Semi-Supervised Learning

In step 4, the ensemble model was applied to the initial labelled data to predict unlabelled data. Based on the model prediction, the least confident instances were selected for human experts' manual labels and then added to the training data. At the same time, based on the model accuracy and label confidence trade-off (Rizve et al., 2021), instances with pseudo-label confidence larger than 0.8 were added to the training data as well.

After several iterations of the labelling and model-updating processes, the model accuracy could not be improved, and we did not do any further iterations. Figure 7 shows the fully labelled data in the two-dimensional space using the first two principal components of the code space.

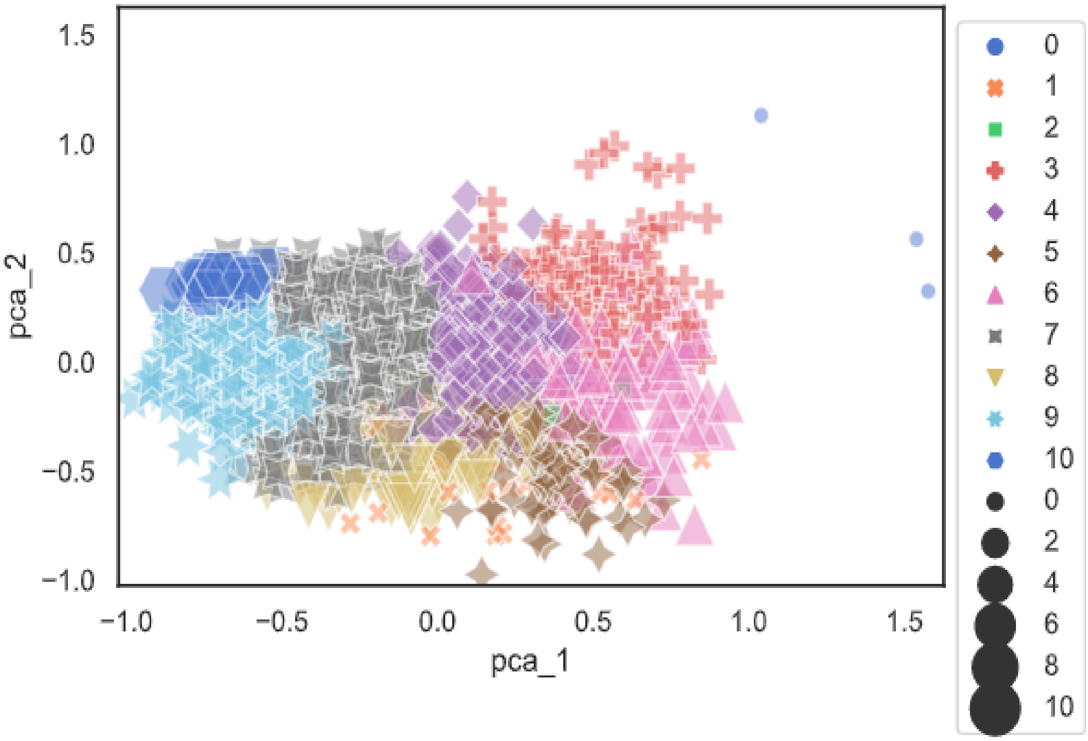


Figure 7. Fully labelled data in the two-dimensional space in step 4.

Note that, in the entire labelling process, about 5% of the instances in the studied sample were manually labelled by human experts. The last column in Table 3 shows the proportion of each group in the process profiles in the fully labelled data. The labels in Table 3 largely align with the raw scores on the booklet from low (label 0) to high (label 11). There is a very small number of students at both ends (disengaged group and the highest-performing groups) and a relatively large number of students (Process Profile 4 and Process Profile 7) in the middle of the profiles. Overall, about 65% of students in the studied sample are in the low- or very-low-score profiles, and 35% are in the medium- or higher-score profiles, which largely agrees with the 2017 national achievement–level results for eighth grade mathematics (that is, 30% below basic, 36% basic, 24% proficient, 10% advanced), even though we did not use the reported scores in our analysis. The process profiles in Table 3 provide more context about the test-taking process and engagement than the performance scores alone.

6. Discussion

Digital-based and technology-enhanced large-scale assessments provide much richer data than traditional assessments, and these richer data offer opportunities to understand students' test-taking and problem-solving processes and to pursue assessment innovations to better serve learning and teaching. In this paper, as a proof of concept, we analyzed the released 2017 NAEP Grade 8 math data, which is about 2% of the total participating students in the 2017 NAEP Grade 8 Mathematics Assessment. The study shows the promise of the powerful AI tools in deepening our understanding at scale of students' entire test-taking processes and performance with context, to address questions that may be related to students' affective states, motivation, and strategies beyond a performance score.

Based on students' interaction with the assessment platform and their responses to the items, we used an HAI approach to analyze the multi-source data (response, time, tool use) and created about a dozen preliminary process profiles for the studied student sample. In the HAI framework, human knowledge, judgment, and decision are reflected in each step in building the AI models. Content knowledge is at the centre of the analysis process to make critical decisions to discover knowledge, create profiles, and annotate instances that are challenging to ML models. The role of ML tools is to help human experts annotate the complex data effectively and efficiently and then generalize their annotations to less-challenging instances. Our study shows the potential of HAI approaches in maximizing AI power and minimizing redundancy of human labour in process data analysis of large-scale assessments.

As discussed in many studies on assessment data (Ercikan & Pellegrino, 2017; Ercikan et al., 2020; Guo, 2022; Guo & Ercikan, 2021a; Pohl et al., 2021; Rios & Guo, 2020), test-taking behaviours are not a nuisance factor that may confound measurement but an aspect that provides important information on how test-takers approach tasks. It has been observed that test-taking behaviours have different relationships with performance for test-taker groups with different pre-knowledge, backgrounds, social-cultural norms, and learning and assessment experiences, which may lead to problematic comparison of performance and difficulty in score interpretation if not considered. Compared to the aforementioned assessment studies that use static process data features or behaviours on an individual task, our exploration showed that sequential process data analysis, integrated with responses and scores, on the entire assessment could provide a rich and holistic picture of students' performance and test-taking behaviours. For example, in the studied sample, far more process profiles were associated with low scores than with high scores, which could help educators learn more about their students, potentially starting a conversation, and provide clues for them to prepare different strategies for teaching and learning intervention in their classrooms or schools. For example, in Profile 0, even though a very small portion (0.27%), students are so unique that educators need to first figure out whether they had the knowledge but just did not want to try, or they did not have the knowledge. For Profile 1, students not only need to learn content, but also need to try harder on the second item block and regulate tool uses. Students in both Profiles 2 and 3 need to learn content, but the data show that students in Profile 2 were likely distracted by tools; they need to either get familiar with digital tools or regulate their tool use. Such process data research could help educators and researchers move one step closer to being able to decide whether a low score was caused by lack of test motivation, use of RG strategies with or without time pressure, test speededness, lack of knowledge, etc.

It is worth noting that some of the preliminary profiles observed in the current study in the assessment environment are similar to those observed in the learning environment. For example, on the studied assessment, low-performing students in Profiles 5 and 6 were potentially struggling by spending prolonged time on one or a few problems, while the high-performing students in Profiles 10 and 11 managed their test-taking process better, as was observed in the learning environment (R. Baker et al., 2010; R. Baker, 2021; Lagud & Rodrigo, 2010; Nawaz et al., 2020). However, in the assessment environment, the timed working session puts pressure on a student, and test speededness is a concern, particularly for low-performing students who may need more time to process and solve a problem. Unlike in a learning environment where students may game the system by abusing hints, in the assessment environment, students used mixed strategies when solving different problems, including the rapid-responding strategy, in order to finish all problems in the session.

6.1 Potential Uses

There are two major potential uses of such HAI methodologies in large-scale assessments: (1) provide a richer picture beyond what students know and can do to educators and policy makers, and (2) help to improve assessment design and tasks.

6.1.1 Potential Uses to Influence Policies

Large-scale, low-stakes assessments, such as NAEP and PISA, have a sophisticated survey design, so that the samples that take the different booklets are randomly equivalent. The current study purposely created comparable process data features on one released booklet, so that similar steps in the architecture (refer to Figure 1) could be applied to the remaining booklets. Use of the initially labelled data, as was done in the current study, as the starting point in the active learning step would speed up the annotation process and reduce computation demand and thus create process profiles for individual students in the total sample.

Reporting a process profile, including both performance scores and process data as explored in the current study, may help provide context to the assessment results, which could lead to a richer comparison across groups by acknowledging heterogeneity within and across student groups, a better understanding and more valid interpretation of performance, and more informed decision making for educational policies to improve students' learning experiences and educational outcomes. Such process profiles could also complement NAEP questionnaire data from students, teachers, and schools to influence educational policies.

6.1.2 Potential Uses for Assessment Innovation and Task Design

Currently, a lot of large-scale assessments are administered digitally, which offers opportunities for assessment innovation (such as using interactive and simulation tasks) to measure complex skills such as creative thinking and problem solving for providing feedback to learners and teachers (International Test Commission and Association of Test Publishers, 2022; National Research Council, 2000; National Center for Education Statistics, 2022; Organisation for Economic Co-operation and Development, 2020). Analyzing process data at either item level or test level can help create meaningful process features, indicators, and process profiles, with or without a psychometric latent model, to provide information about students' problem-solving behaviours or test-taking processes that is richer than a single performance score. This information can in turn help improve current digital-based assessment designs to be more engaging, effective, and valid in measuring students' knowledge and skills. For a simple example, if students show repeated RG behaviours from process data, a flag could be implemented in the system to encourage their engagement with the assessment (Wise, 2021). Similarly, if a student spent too much time on one question, a reminder could be implemented to suggest that the student move on to the next question and come back to it later if time permits.

The HAI methodologies could be used in studies to improve assessment design, for example, focusing on digital tool uses to evaluate which digital tools are useful and which are distracting to different student groups, and how to improve them. Moreover, knowledge gained from the digital-based assessments could help to design next-generation assessments and innovative tasks to measure complex constructs and skills that take into account students' response processes when evaluating their skills. HAI approaches, such as the one we explored in the current study, could help make sense of the complex data at scale, including data collected from audio, video, image, text, and log files. Such HAI approaches would be particularly useful in analyzing and interpreting item-level process data, in order to provide insight into innovative task creation to elicit desired response behaviours in the test-taking processes, further helping develop psychometric models for scoring next-generation assessments (Ercikan et al., 2023; Foster & Piacentini, 2023).

6.2 Limitations and Future Directions

As a proof of concept, data from one released NAEP booklet were used in the study, producing about a dozen preliminary process profiles. With more data from other NAEP booklets, the results in the current study could be improved by experimenting with finer discretization of features (more timing and tool use categories, for example) and fine tuning of algorithms, for example. Even though at each step in our study, we evaluated different ways of feature creation, different numbers for the code space dimension, different ML methods, and different random seeds, with humans in the centre for decision making as safeguards, sensitivity analysis is necessary, particularly when data from more booklets are acquired for generalization.

Another limitation of the current study involves the interpretation and validation of the ML predicted (pseudo-) profiles, which need to be addressed by linking the process profiles to the other external data source (such as student background data, indices from survey questionnaires, or teacher evaluations), even though these external data may not contain fine-grained low-level and temporal information and may not be directly associated with the test-taking behaviours on the NAEP mathematics assessment. Further investigations also need to involve educators, who can help to evaluate and improve the preliminary profiles for possible revision, combination, and creation of new profiles, so that such profiles will be truly useful and valid. Note that the results in the study cannot be used to establish a cause-and-effect relationship between test-taking behaviours and student achievement, since different behaviours may be associated with similar performance levels and vice versa (Ercikan & Pellegrino, 2017; Guo, 2022).

In addition, in the current study, the deep learning algorithm autoencoder with LSTM was used to preserve sequential information. It may be worth investigating whether new features can be created to reflect such sequential information to use shallow ML models instead of deep learning models, so that the annotating process is more explainable. Further studies could also investigate how to create new features to study interaction between content areas and test-taking behaviours.

More important, if the assessment is designed to measure both knowledge and problem-solving processes, future research needs to focus on how to expand existing psychometric latent models (Johnson & Liu, 2022; Levy, 2020; van der Linden, 2007; Wang & Xu, 2015) to integrate process data with response data in the scoring models. When AI techniques are used, researchers need to work with stakeholders to identify the specific purposes of studying complex data from large-scale assessments to better inform assessment developers, educators, and policy makers.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgements

The authors would like to thank the editor and anonymous reviewers for their diligent review and insightful comments for improving the quality of the manuscript.

Funding

This project has been funded at least in part with federal funds from the National Center for Education Statistics (NCES) of the Institute of Education Sciences (IES) 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 endorsement of the U.S. Government.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/

Baker, R., D'Mello, S., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

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.

Baker, R., Ogan, A., Madaio, M., & Walker, E. (2019). Culture in computer-based learning systems: Challenges and opportunities. Computer-based learning in context. *Zenodo*. https://doi.org/10.5281/zenodo.4057223

Baker, T., Smith, L., & Anissa, N. (2019). *Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges* (Report). NESTA. London, UK. https://www.nesta.org.uk/report/education-rebooted

Bennett, R. E., Zhang, M., Sinharay, S., Guo, H., & Deane, P. (2022). Are there distinctive profiles in examinee essay-writing processes? *Educational Measurement: Issues and Practice*, *41*(2), 55–69. https://doi.org/10.1111/emip.12469

Biswas, G., Segedy, J., & Bunchongchit, K. (2016). From design to implementation to practice a learning by teaching system: Betty's brain. *International Journal of Artificial Intelligence in Education*, *26*, 350–364. 10.1007/s40593-015-0057-9

Deane, P., Wilson, J., Zhang, M., Li, C., van Rijn, P., Guo, H., Roth, A., Winchester, E., & Richter, T. (2021). The sensitivity of a scenario-based assessment of written argumentation to school differences in curriculum and instruction. *International Journal of Artificial Intelligence in Education*, *31*(1), 57–98. https://doi.org/10.1007/s40593-020-00227-x

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. In N. Foster & M. Piacentini (Eds.), *Innovating assessments to measure and support complex skills* (pp. 211–228). OECD Publishing. https://doi.org/10.1787/7b3123f1-en

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

Foster, N., & Piacentini, M. (Eds.). (2023). *Innovating assessments to measure and support complex skills*. OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Geron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media. https://www.bibsonomy.org/bibtex/2a91270a3a516f4edaa5d459c40317fcc/achakraborty

Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is deep learning the best approach to knowledge tracing? *Journal of Educational Data Mining*, *12*(3), 31–54. https://doi.org/10.5281/zenodo.4143614

Gordon, E. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice*, *39*(3), 72–78. 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. https://doi.org/10.1016/j.chb.2016.02.095

Guo, H. (2022). How did students engage with a remote educational assessment? A case study. *Educational Measurement: Issues and Practice*, *41*(3), 58–68. https://doi.org/10.1111/emip.12476

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., Zhang, M., Deane, P., & Bennett, R. (2020). Effects of scenario-based assessment on students' writing processes. *Journal of Educational Data Mining*, *12*(1), 19–45. https://doi.org/10.5281/zenodo.3911797

International Test Commission and Association of Test Publishers. (2022). *Guidelines for technology-based assessment*. https://www.intestcom.org/news/38

Johnson, M. S., & Liu, X. (2022). *Psychometric considerations for the joint modeling of response and process data* [Paper presented at the 2022 IMPS annual meeting, 11–15 July 2022, Bologna, Italy].

Kleinman, E., Shergadwala, M., Teng, Z., Villareale, J., Bryant, A., Zhu, J., & Seif El-Nasr, M. (2022). Analyzing students' problem-solving sequences: A human-in-the-loop approach. *Journal of Learning Analytics*, *9*(2), 138–160. https://doi.org/10.18608/jla.2022.7465

Lagud, M. C. V., & Rodrigo, M. M. T. (2010). The affective and learning profiles of students using an intelligent tutoring system for algebra. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Proceedings of the 10th International Conference on*

Intelligent Tutoring Systems (ITS 2010), 14–18 June 2010, Pittsburgh, Pennsylvania, USA (pp. 255–263). Springer. https://doi.org/10.1007/978-3-642-13388-6_30

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

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

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

National Center for Education Statistics. (2017). NAEP 2017 sample design. https://nces.ed.gov/nationsreportcard/tdw/sample_design/2017/naep_2017_sample_design.aspx

National Center for Education Statistics. (2021). NAEP DBA tutorial. https://enaep-public.naepims.org/2017/EN/main.html?subject=Math8

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

National Research Council. (2000). How people learn: Brain, mind, experience, and school: Expanded edition. https://doi.org/10.17226/9853

Nawaz, S., Kennedy, G., Bailey, J., & Mead, C. (2020). Moments of confusion in simulation-based learning environments. *Journal of Learning Analytics*, *7*(3), 118–137. https://doi.org/10.18608/jla.2020.73.9

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

Organisation for Economic Co-operation and Development. (2020). PISA 2018 database. https://www.oecd.org/pisa/data/2018database/

Paquette, L., & Baker, R. S. (2019). Comparing machine learning to knowledge engineering for student behavior modeling: A case study in gaming the system. *Interactive Learning Environments*, *27*(5-6), 585–597. https://doi.org/10.1080/10494820.2019.1610450

Pellegrino, J. W. (2020). Important considerations for assessment to function in the service of education. *Educational Measurement: Issues and Practice*, *39*(3), 81–85. 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. https://www.science.org/doi/abs/10.1126/science.abd3300

Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: Evidence from the English version of the PISA 2015 science test. *Large-scale Assessments in Education*, *9*(10). https://doi.org/10.1186/s40536-021-00104-6

Radwan, A. M. (2019). Human active learning. In S. M. Brito (Ed.), *Active learning*. IntechOpen. 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/10.1080/15305058.2016.1231193

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 *Proceedings of the 2021 International Conference on Learning Representations* (ICLR 2021), 4 May 2021, Vienna, Austria. https://openreview.net/forum?id=-ODN6SbiUU

Trinh, H. D., Zeydan, E., Giupponi, L., & Dini, P. (2019). Detecting mobile traffic anomalies through physical control channel fingerprinting: A deep semi-supervised approach. *IEEE Access*, *7*, 152187–152201. https://doi.org/10.1109/ACCESS.2019.2947742

Ulitzsch, E., He, Q., & Pohl, S. (2022). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics*, *47*(1), 3–35. https://doi.org/10.3102/10769986211010467

van der Linden. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. https://doi.org/10.1111/bmsp.12054

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

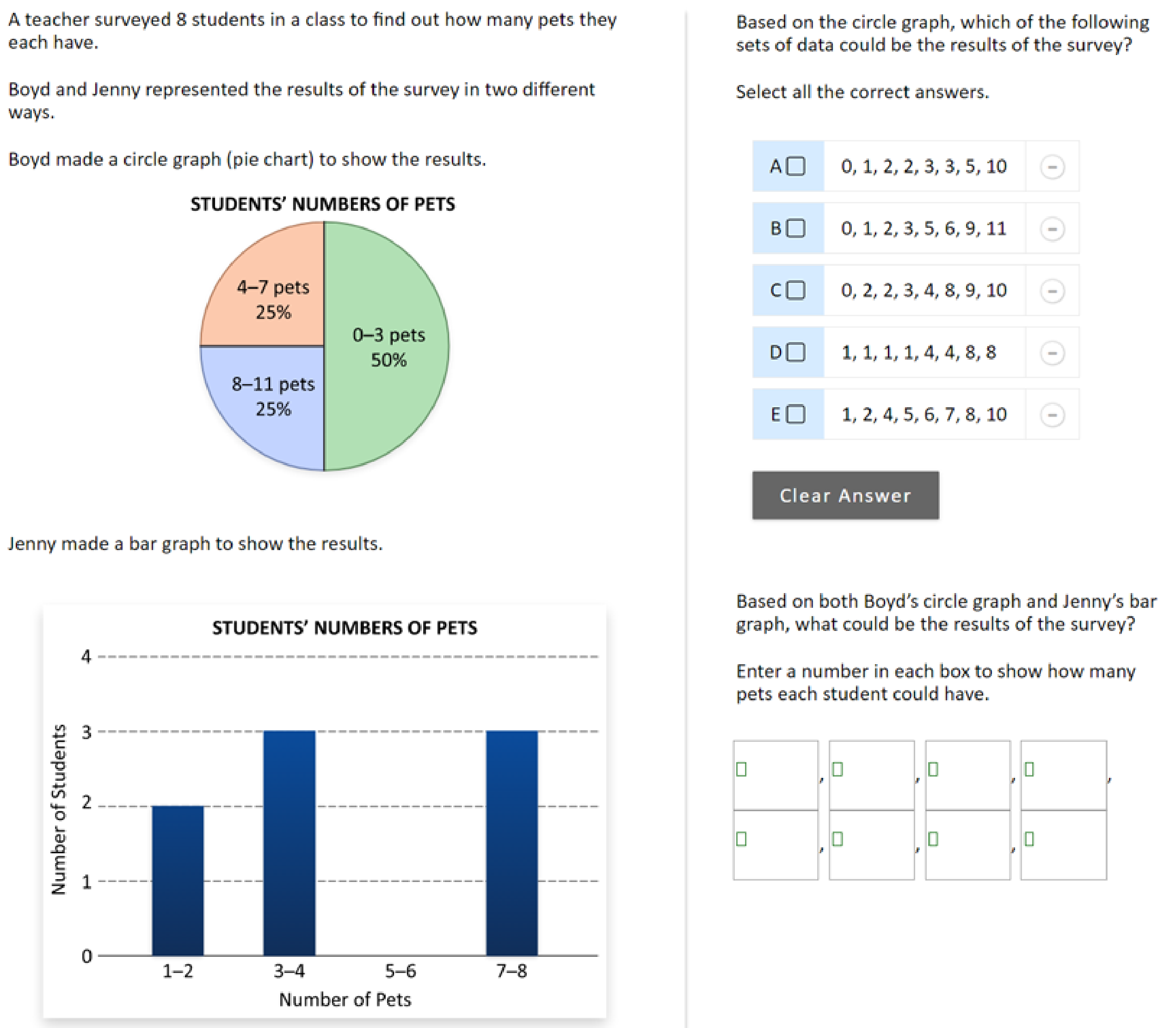
Wise, S. (2021). Six insights regarding test-taking disengagement. *Educational Research and Evaluation*, *26*(5-6), 328–338. https://doi.org/10.1080/13803611.2021.1963942

Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. V. (2020). Unsupervised data augmentation for consistency training. In *Proceedings of the 34th Conference on Neural Information Processing Systems* (NeurIPS 2020), 6–12 December 2020, Vancouver, British Columbia, Canada (pp. 6256–6268). Curran Associates. https://dl.acm.org/doi/pdf/10.5555/3495724.3496249

Zhu, X., Lafferty, J., & Ghahramani, Z. (2003). Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the ICML 2003 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining,* 21 August 2003, Washington, DC, USA (pp. 58–65). Springer. https://doi.org/10.1007/978-3-319-12637-1_27

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*. OECD Publishing. https://doi.org/10.1787/9789264273955-en

Appendix



A teacher surveyed 8 students in a class to find out how many pets they each have.

Boyd and Jenny represented the results of the survey in two different ways.

Boyd made a circle graph (pie chart) to show the results.

Jenny made a bar graph to show the results.

Based on the circle graph, which of the following sets of data could be the results of the survey?

Select all the correct answers.

A □	0, 1, 2, 2, 3, 3, 5, 10	⊖
B □	0, 1, 2, 3, 5, 6, 9, 11	⊖
C □	0, 2, 2, 3, 4, 8, 9, 10	⊖
D □	1, 1, 1, 1, 4, 4, 8, 8	⊖
E □	1, 2, 4, 5, 6, 7, 8, 10	⊖

Based on both Boyd's circle graph and Jenny's bar graph, what could be the results of the survey?

Enter a number in each box to show how many pets each student could have.

□ , □ , □ , □ ,
□ , □ , □ , □

Figure 8. Item 13 (released) in block 8M3.

Table A.1. A representative instance from Profile 0: Unengaged group. This student worked on two items only.

	Time_M3 (min)	Time_M9 (min)	Total Score	Total Tool Use
Rep	0.0	0.90	0.0	3.00
Clust_mean	0.0	7.56	1.0	3.50
Clust_std	0.0	10.30	1.0	2.29

Score	Time	Tool	Raw_Time	Raw_Tool	Item#	Block	Difficulty	Type	Content_short
0	4	3	46	3	1	8M9	Medium	MC	Measurement
0	1	0	5	0	2	8M9	Easy	MC	Geometry
0	0	0	0	0	3	8M9	Medium	MC	Numbers
0	0	0	0	0	4	8M9	Easy	SR	Numbers
0	0	0	0	0	5	8M9	Hard	MC	Data and Stats
0	0	0	0	0	6	8M9	Hard	MC	Geometry
0	0	0	0	0	7	8M9	Easy	SCR	Algebra
0	0	0	0	0	8	8M9	Hard	MC	Measurement
0	0	0	0	0	9	8M9	Hard	SCR	Geometry
0	0	0	0	0	10	8M9	Hard	MC	Numbers
0	0	0	0	0	11	8M9	Medium	MC	Algebra
0	0	0	0	0	12	8M9	Hard	SCR	Numbers
0	0	0	0	0	13	8M9	Hard	SCR	Algebra
0	0	0	0	0	14	8M9	Hard	MC	Data and Stats
0	0	0	0	0	15	8M9	Medium	MC	Algebra
0	0	0	0	0	16	8M9	Hard	MC	Numbers
0	0	0	0	0	17	8M9	Medium	MC	Algebra
0	0	0	0	0	18	8M9	Hard	MC	Data and Stats
0	0	0	0	0	19	8M9	Easy	MC	Algebra
0	0	0	0	0	1	8M3	Easy	MC	Numbers
0	0	0	0	0	2	8M3	Easy	SR	Data and Stats
0	0	0	0	0	3	8M3	Medium	SCR	Numbers
0	0	0	0	0	4	8M3	Easy	SR	Algebra
0	0	0	0	0	5	8M3	Easy	SR	Measurement
0	0	0	0	0	6	8M3	Medium	SR	Algebra
0	0	0	0	0	7	8M3	Hard	SCR	Measurement
0	0	0	0	0	8	8M3	Hard	MC	Geometry
0	0	0	0	0	9	8M3	Easy	SR	Measurement
0	0	0	0	0	10	8M3	Hard	SR	Geometry
0	0	0	0	0	11	8M3	Hard	SCR	Algebra
0	0	0	0	0	12	8M3	Hard	SR	Geometry
0	0	0	0	0	13	8M3	Hard	ECR	Data and Stats
0	0	0	0	0	14	8M3	Hard	SR	Data and Stats
0	0	0	0	0	15	8M3	Hard	SCR	Geometry

Note: In the item type column (Type), MC = multiple choice, SR = composite selected response, SCR = short constructed response, and ECR = extended constructed response.

Table A.2. A representative instance from Profile 1: Low-performing and disengaged group, with some tool play. This student used relatively more tools than the whole sample, particularly on block 8M9. This student became less engaged with less than 20 minutes left on block 8M3.

	Time_M3 (min)	**Time_M9 (min)**	**Total Score**	**Total Tool Use**
Rep	18.57	26.08	9.00	42.00
Clust_mean	15.18	26.63	8.23	79.39
clust_std	9.46	4.63	3.30	42.42

Score	Time	Tool	Raw_Time	Raw_Tool	Item#	Block	Difficulty	Type	Content_short
0	2	2	22	1	1	8M9	Medium	MC	Measurement
0	2	2	23	1	2	8M9	Easy	MC	Geometry
1	2	2	62	3	3	8M9	Medium	MC	Numbers
0	4	1	254	1	4	8M9	Easy	SR	Numbers
0	3	1	117	1	5	8M9	Hard	MC	Data and Stats
1	3	2	66	1	6	8M9	Hard	MC	Geometry
0	3	2	151	1	7	8M9	Easy	SCR	Algebra
1	3	3	112	4	8	8M9	Hard	MC	Measurement
0	2	0	24	0	9	8M9	Hard	SCR	Geometry
0	2	2	58	2	10	8M9	Hard	MC	Numbers
0	3	2	88	1	11	8M9	Medium	MC	Algebra
0	3	2	104	1	12	8M9	Hard	SCR	Numbers
0	3	2	104	2	13	8M9	Hard	SCR	Algebra
0	4	3	100	8	14	8M9	Hard	MC	Data and Stats
0	3	3	54	1	15	8M9	Medium	MC	Algebra
0	2	3	22	1	16	8M9	Hard	MC	Numbers
0	2	3	13	2	17	8M9	Medium	MC	Algebra
1	4	3	74	2	18	8M9	Hard	MC	Data and Stats
1	4	3	96	2	19	8M9	Easy	MC	Algebra
0	3	3	35	1	1	8M3	Easy	MC	Numbers
1	4	0	95	0	2	8M3	Easy	SR	Data and Stats
0	2	2	66	1	3	8M3	Medium	SCR	Numbers
2	3	0	81	0	4	8M3	Easy	SR	Algebra
1	3	0	82	0	5	8M3	Easy	SR	Measurement
0	3	0	94	0	6	8M3	Medium	SR	Algebra
0	2	0	27	0	7	8M3	Hard	SCR	Measurement
0	2	0	17	0	8	8M3	Hard	MC	Geometry
0	3	0	63	0	9	8M3	Easy	SR	Measurement
0	3	0	76	0	10	8M3	Hard	SR	Geometry
0	2	0	78	0	11	8M3	Hard	SCR	Algebra
0	4	3	100	2	12	8M3	Hard	SR	Geometry
0	2	0	57	0	13	8M3	Hard	ECR	Data and Stats
0	2	0	46	0	14	8M3	Hard	SR	Data and Stats
0	4	3	185	3	15	8M3	Hard	SCR	Geometry

Note: In the item type column (Type), MC = multiple choice, SR = composite selected response, SCR = short constructed response, and ECR = extended constructed response.

Table A.3. A representative instance from Profile 2: Low performing, with regulated time but potentially distracted by tools. This student had a lot of tool uses.

	Time_M3 (min)	**Time_M9 (min)**	**Total Score**	**Total Tool Use**
Rep	28.87	25.33	13.00	209.00
Clust_mean	24.38	28.17	10.56	218.22
clust_std	3.89	2.95	3.80	62.44

Score	Time	Tool	Raw_Time	Raw_Tool	Item#	Block	Difficulty	Type	Content_short
0	2	0	17	0	1	8M9	Medium	MC	Measurement
0	3	2	86	1	2	8M9	Easy	MC	Geometry
0	3	3	99	11	3	8M9	Medium	MC	Numbers
1	2	1	91	1	4	8M9	Easy	SR	Numbers
0	2	3	81	6	5	8M9	Hard	MC	Data and Stats
1	2	2	33	2	6	8M9	Hard	MC	Geometry
2	2	3	107	7	7	8M9	Easy	SCR	Algebra
0	3	3	98	6	8	8M9	Hard	MC	Measurement
1	4	3	125	11	9	8M9	Hard	SCR	Geometry
0	4	3	125	6	10	8M9	Hard	MC	Numbers
0	3	3	72	11	11	8M9	Medium	MC	Algebra
0	3	3	107	5	12	8M9	Hard	SCR	Numbers
0	2	3	97	4	13	8M9	Hard	SCR	Algebra
0	3	3	81	7	14	8M9	Hard	MC	Data and Stats
1	4	3	61	6	15	8M9	Medium	MC	Algebra
1	3	3	31	2	16	8M9	Hard	MC	Numbers
0	4	3	59	6	17	8M9	Medium	MC	Algebra
0	4	3	80	5	18	8M9	Hard	MC	Data and Stats
0	3	3	57	2	19	8M9	Easy	MC	Algebra
0	3	3	34	5	1	8M3	Easy	MC	Numbers
1	4	3	116	14	2	8M3	Easy	SR	Data and Stats
0	4	2	134	1	3	8M3	Medium	SCR	Numbers
2	4	3	116	11	4	8M3	Easy	SR	Algebra
1	3	3	118	7	5	8M3	Easy	SR	Measurement
1	4	3	133	11	6	8M3	Medium	SR	Algebra
0	4	3	135	6	7	8M3	Hard	SCR	Measurement
0	3	3	63	4	8	8M3	Hard	MC	Geometry
0	3	3	78	7	9	8M3	Easy	SR	Measurement
1	4	3	109	7	10	8M3	Hard	SR	Geometry
0	4	3	221	9	11	8M3	Hard	SCR	Algebra
0	4	3	127	5	12	8M3	Hard	SR	Geometry
0	3	3	171	9	13	8M3	Hard	ECR	Data and Stats
0	4	3	110	4	14	8M3	Hard	SR	Data and Stats
0	2	3	57	10	15	8M3	Hard	SCR	Geometry

Note: In the item type column (Type), MC = multiple choice, SR = composite selected response, SCR = short constructed response, and ECR = extended constructed response.

Table A.4. A representative instance from Profile 3: Very low performing but overall engaged, with regulated time and tool uses. This student may have used mixed strategies by spending prolonged time (Time = 5) on some items and relatively short time (Time = 2) on some other items.

	Time_M3 (min)	**Time_M9 (min)**	**Total Score**	**Total Tool Use**
Rep	24.25	27.20	2.00	13.00
Clust_mean	14.24	18.77	5.62	19.04
clust_std	6.62	8.33	2.68	15.80

Score	Time	Tool	Raw_Time	Raw_Tool	Item#	Block	Difficulty	Type	Content_short
0	2	0	16	0	1	8M9	Medium	MC	Measurement
0	3	2	82	1	2	8M9	Easy	MC	Geometry
0	3	1	84	1	3	8M9	Medium	MC	Numbers
0	3	0	130	0	4	8M9	Easy	SR	Numbers
0	2	1	83	1	5	8M9	Hard	MC	Data and Stats
0	2	0	32	0	6	8M9	Hard	MC	Geometry
0	3	2	140	1	7	8M9	Easy	SCR	Algebra
0	2	1	65	1	8	8M9	Hard	MC	Measurement
0	2	0	56	0	9	8M9	Hard	SCR	Geometry
0	3	1	116	1	10	8M9	Hard	MC	Numbers
0	3	3	79	3	11	8M9	Medium	MC	Algebra
1	4	0	152	0	12	8M9	Hard	SCR	Numbers
0	2	0	61	0	13	8M9	Hard	SCR	Algebra
0	2	2	30	1	14	8M9	Hard	MC	Data and Stats
0	2	0	37	0	15	8M9	Medium	MC	Algebra
0	4	0	41	0	16	8M9	Hard	MC	Numbers
0	5	0	97	0	17	8M9	Medium	MC	Algebra
0	5	0	230	0	18	8M9	Hard	MC	Data and Stats
0	3	0	49	0	19	8M9	Easy	MC	Algebra
0	3	0	41	0	1	8M3	Easy	MC	Numbers
1	2	0	49	0	2	8M3	Easy	SR	Data and Stats
0	4	2	132	2	3	8M3	Medium	SCR	Numbers
0	2	0	56	0	4	8M3	Easy	SR	Algebra
0	2	0	50	0	5	8M3	Easy	SR	Measurement
0	3	0	119	0	6	8M3	Medium	SR	Algebra
0	3	2	100	1	7	8M3	Hard	SCR	Measurement
0	4	0	121	0	8	8M3	Hard	MC	Geometry
0	5	0	197	0	9	8M3	Easy	SR	Measurement
0	4	0	147	0	10	8M3	Hard	SR	Geometry
0	2	0	81	0	11	8M3	Hard	SCR	Algebra
0	2	0	29	0	12	8M3	Hard	SR	Geometry
0	2	0	139	0	13	8M3	Hard	ECR	Data and Stats
0	4	0	140	0	14	8M3	Hard	SR	Data and Stats
0	2	0	32	0	15	8M3	Hard	SCR	Geometry

Note: In the item type column (Type), MC = multiple choice, SR = composite selected response, SCR = short constructed response, and ECR = extended constructed response.

Table A.5. A representative instance from Profile 4: Low performing and disengaged on the second item block, with regulated tool uses. The student spent full time on block 8M9, but less than 20 minutes on block 8M3. Used the mixed strategies of rapid responding and prolonged duration times.

	Time_M3 (min)	**Time_M9 (min)**	**Total Score**	**Total Tool Use**
Rep	19.02	30.30	10.00	34.00
Clust_mean	21.99	25.89	11.39	52.58
clust_std	6.06	4.32	2.57	44.04

Score	**Time**	**Tool**	**Raw_Time**	**Raw_Tool**	**Item#**	**Block**	**Difficulty**	**Type**	**Content_short**
1	2	0	15	0	1	8M9	Medium	MC	Measurement
0	2	2	56	2	2	8M9	Easy	MC	Geometry
0	5	3	271	4	3	8M9	Medium	MC	Numbers
0	4	3	244	3	4	8M9	Easy	SR	Numbers
0	3	2	131	3	5	8M9	Hard	MC	Data and Stats
0	2	0	42	0	6	8M9	Hard	MC	Geometry
0	4	3	208	3	7	8M9	Easy	SCR	Algebra
0	3	3	132	3	8	8M9	Hard	MC	Measurement
0	2	2	56	2	9	8M9	Hard	SCR	Geometry
0	3	2	115	2	10	8M9	Hard	MC	Numbers
1	2	0	62	0	11	8M9	Medium	MC	Algebra
0	1	0	1	0	12	8M9	Hard	SCR	Numbers
0	3	2	126	2	13	8M9	Hard	SCR	Algebra
0	3	3	74	2	14	8M9	Hard	MC	Data and Stats
1	2	0	24	0	15	8M9	Medium	MC	Algebra
0	2	0	14	0	16	8M9	Hard	MC	Numbers
1	2	0	21	0	17	8M9	Medium	MC	Algebra
0	2	0	44	0	18	8M9	Hard	MC	Data and Stats
1	5	3	171	2	19	8M9	Easy	MC	Algebra
0	2	0	15	0	1	8M3	Easy	MC	Numbers
1	3	0	57	0	2	8M3	Easy	SR	Data and Stats
0	2	2	60	2	3	8M3	Medium	SCR	Numbers
1	3	0	61	0	4	8M3	Easy	SR	Algebra
1	3	0	108	0	5	8M3	Easy	SR	Measurement
0	3	0	89	0	6	8M3	Medium	SR	Algebra
0	4	3	138	2	7	8M3	Hard	SCR	Measurement
1	3	0	52	0	8	8M3	Hard	MC	Geometry
1	4	0	88	0	9	8M3	Easy	SR	Measurement
0	2	0	57	0	10	8M3	Hard	SR	Geometry
0	2	0	68	0	11	8M3	Hard	SCR	Algebra
0	4	3	106	2	12	8M3	Hard	SR	Geometry
0	2	0	141	0	13	8M3	Hard	ECR	Data and Stats
0	2	0	62	0	14	8M3	Hard	SR	Data and Stats
0	1	0	17	0	15	8M3	Hard	SCR	Geometry

Note: In the item type column (Type), MC = multiple choice, SR = composite selected response, SCR = short constructed response, and ECR = extended constructed response.

Table A.6. A representative instance from Profile 5: Low performing, potentially speeded, and struggling, with moderate to high tool uses. This student spent a relatively long time in the first half of block 8M9 (prolonged times on a few items) and did not reach the last few items. However, this student managed to finish block 8M3, but with too many tool uses on the latter-half items.

	Time_M3 (min)	Time_M9 (min)	Total Score	Total Tool Use
Rep	24.70	30.22	13.00	127.00
Clust_mean	28.42	30.28	10.68	135.95
clust_std	2.73	0.10	3.43	75.81

Score	Time	Tool	Raw_Time	Raw_Tool	Item#	Block	Difficulty	Type	Content_short
0	4	2	54	2	1	8M9	Medium	MC	Measurement
1	3	2	78	1	2	8M9	Easy	MC	Geometry
0	5	3	395	5	3	8M9	Medium	MC	Numbers
1	4	1	225	1	4	8M9	Easy	SR	Numbers
0	2	3	103	5	5	8M9	Hard	MC	Data and Stats
0	3	3	64	7	6	8M9	Hard	MC	Geometry
2	2	3	82	7	7	8M9	Easy	SCR	Algebra
0	4	1	210	1	8	8M9	Hard	MC	Measurement
0	3	3	105	8	9	8M9	Hard	SCR	Geometry
1	5	1	219	1	10	8M9	Hard	MC	Numbers
0	2	2	50	1	11	8M9	Medium	MC	Algebra
0	4	2	177	1	12	8M9	Hard	SCR	Numbers
0	1	0	14	0	13	8M9	Hard	SCR	Algebra
0	0	0	0	0	14	8M9	Hard	MC	Data and Stats
0	0	0	0	0	15	8M9	Medium	MC	Algebra
0	0	0	0	0	16	8M9	Hard	MC	Numbers
0	0	0	0	0	17	8M9	Medium	MC	Algebra
0	0	0	0	0	18	8M9	Hard	MC	Data and Stats
0	0	0	0	0	19	8M9	Easy	MC	Algebra
0	4	0	74	0	1	8M3	Easy	MC	Numbers
1	2	0	45	0	2	8M3	Easy	SR	Data and Stats
0	2	0	65	0	3	8M3	Medium	SCR	Numbers
0	3	0	83	0	4	8M3	Easy	SR	Algebra
2	2	0	78	0	5	8M3	Easy	SR	Measurement
1	4	3	145	8	6	8M3	Medium	SR	Algebra
0	2	3	46	8	7	8M3	Hard	SCR	Measurement
1	4	3	79	4	8	8M3	Hard	MC	Geometry
1	4	3	115	13	9	8M3	Easy	SR	Measurement
0	4	0	100	0	10	8M3	Hard	SR	Geometry
0	2	3	116	7	11	8M3	Hard	SCR	Algebra
0	3	3	84	4	12	8M3	Hard	SR	Geometry
2	3	3	211	37	13	8M3	Hard	ECR	Data and Stats
0	4	3	121	6	14	8M3	Hard	SR	Data and Stats
0	3	0	111	0	15	8M3	Hard	SCR	Geometry

Note: In the item type column (Type), MC = multiple choice, SR = composite selected response, SCR = short constructed response, and ECR = extended constructed response.

Table A.7. A representative instance from Profile 6: Low performing, potentially speeded, and struggling, with low tool uses. This student used mixed strategies, spent a relatively long time on earlier items in block 8M9 (prolonged times with Time = 5), and did not get to the last few items. However, this student managed to finish block 8M3 with rapid responding.

	Time_M3 (min)	**Time_M9 (min)**	**Total Score**	**Total Tool Use**
Rep	30.30	30.20	9.00	14.00
Clust_mean	25.63	29.72	7.33	15.78
clust_std	5.99	2.33	2.59	6.80

Score	Time	Tool	Raw_Time	Raw_Tool	Item#	Block	Difficulty	Type	Content_short
0	2	3	18	3	1	8M9	Medium	MC	Measurement
1	4	2	136	1	2	8M9	Easy	MC	Geometry
1	5	3	295	4	3	8M9	Medium	MC	Numbers
1	4	1	205	1	4	8M9	Easy	SR	Numbers
0	2	0	66	0	5	8M9	Hard	MC	Data and Stats
1	2	0	50	0	6	8M9	Hard	MC	Geometry
2	5	2	314	1	7	8M9	Easy	SCR	Algebra
0	4	1	233	1	8	8M9	Hard	MC	Measurement
0	3	2	97	1	9	8M9	Hard	SCR	Geometry
0	4	1	139	1	10	8M9	Hard	MC	Numbers
0	4	2	100	1	11	8M9	Medium	MC	Algebra
0	4	0	127	0	12	8M9	Hard	SCR	Numbers
0	1	0	18	0	13	8M9	Hard	SCR	Algebra
0	1	0	10	0	14	8M9	Hard	MC	Data and Stats
0	0	0	0	0	15	8M9	Medium	MC	Algebra
0	0	0	0	0	16	8M9	Hard	MC	Numbers
0	0	0	0	0	17	8M9	Medium	MC	Algebra
0	0	0	0	0	18	8M9	Hard	MC	Data and Stats
0	0	0	0	0	19	8M9	Easy	MC	Algebra
0	3	0	42	0	1	8M3	Easy	MC	Numbers
1	4	0	79	0	2	8M3	Easy	SR	Data and Stats
0	3	0	78	0	3	8M3	Medium	SCR	Numbers
0	4	0	118	0	4	8M3	Easy	SR	Algebra
2	5	0	221	0	5	8M3	Easy	SR	Measurement
0	4	0	191	0	6	8M3	Medium	SR	Algebra
0	3	0	120	0	7	8M3	Hard	SCR	Measurement
0	4	0	121	0	8	8M3	Hard	MC	Geometry
0	4	0	146	0	9	8M3	Easy	SR	Measurement
0	5	0	187	0	10	8M3	Hard	SR	Geometry
0	3	0	196	0	11	8M3	Hard	SCR	Algebra
0	5	0	173	0	12	8M3	Hard	SR	Geometry
0	2	0	116	0	13	8M3	Hard	ECR	Data and Stats
0	2	0	18	0	14	8M3	Hard	SR	Data and Stats
0	1	0	8	0	15	8M3	Hard	SCR	Geometry

Note: In the item type column (Type), MC = multiple choice, SR = composite selected response, SCR = short constructed response, and ECR = extended constructed response.

Table A.8. A representative instance from Profile 7: Medium-performing group in all dimensions. This student had overall average scores and regulated time and tool uses.

	Time_M3 (min)	Time_M9 (min)	Total Score	Total Tool Use
Rep	29.25	22.60	26.00	35.00
Clust_mean	23.74	26.63	22.83	37.56
clust_std	4.69	3.27	3.93	12.60

Score	Time	Tool	Raw_Time	Raw_Tool	Item#	Block	Difficulty	Type	Content_short
1	3	2	33	2	1	8M9	Medium	MC	Measurement
1	2	0	54	0	2	8M9	Easy	MC	Geometry
1	2	0	57	0	3	8M9	Medium	MC	Numbers
1	2	0	95	0	4	8M9	Easy	SR	Numbers
0	2	0	52	0	5	8M9	Hard	MC	Data and Stats
1	2	2	37	2	6	8M9	Hard	MC	Geometry
2	2	0	67	0	7	8M9	Easy	SCR	Algebra
0	3	3	95	2	8	8M9	Hard	MC	Measurement
0	4	0	119	0	9	8M9	Hard	SCR	Geometry
1	3	1	81	1	10	8M9	Hard	MC	Numbers
0	2	3	67	5	11	8M9	Medium	MC	Algebra
2	3	0	124	0	12	8M9	Hard	SCR	Numbers
1	3	0	100	0	13	8M9	Hard	SCR	Algebra
0	3	0	89	0	14	8M9	Hard	MC	Data and Stats
1	2	0	39	0	15	8M9	Medium	MC	Algebra
1	5	3	91	7	16	8M9	Hard	MC	Numbers
0	2	0	15	0	17	8M9	Medium	MC	Algebra
0	3	0	52	0	18	8M9	Hard	MC	xxx
0	3	3	51	2	19	8M9	Easy	MC	Algebra
0	4	0	46	0	1	8M3	Easy	MC	Numbers
1	5	3	184	4	2	8M3	Easy	SR	Data and Stats
1	4	3	140	3	3	8M3	Medium	SCR	Numbers
2	3	3	57	2	4	8M3	Easy	SR	Algebra
2	4	3	142	2	5	8M3	Easy	SR	Measurement
1	4	0	153	0	6	8M3	Medium	SR	Algebra
0	4	3	182	3	7	8M3	Hard	SCR	Measurement
0	2	0	42	0	8	8M3	Hard	MC	Geometry
1	3	0	74	0	9	8M3	Easy	SR	Measurement
2	3	0	79	0	10	8M3	Hard	SR	Geometry
2	3	0	213	0	11	8M3	Hard	SCR	Algebra
0	3	0	86	0	12	8M3	Hard	SR	Geometry
1	2	0	146	0	13	8M3	Hard	ECR	Data and Stats
0	3	0	90	0	14	8M3	Hard	SR	Data and Stats
0	3	0	79	0	15	8M3	Hard	SCR	Geometry

Note: In the item type column (Type), MC = multiple choice, SR = composite selected response, SCR = short constructed response, and ECR = extended constructed response.

Table A.9. A representative instance from Profile 8: Medium performing and potentially speeded, with high tool use. This student spent prolonged time on some early items and did not get to the last few items in block 8M9. The student managed time better on block 8M3.

	Time_M3 (min)	**Time_M9 (min)**	**Total Score**	**Total Tool Use**
Rep	30.27	30.30	21.00	50.00
Clust_mean	28.74	30.28	19.97	37.78
clust_std	2.80	0.14	3.93	28.06

Score	Time	Tool	Raw_Time	Raw_Tool	Item#	Block	Difficulty	Type	Content_short
0	3	0	32	0	1	8M9	Medium	MC	Measurement
1	2	2	40	1	2	8M9	Easy	MC	Geometry
0	4	1	199	1	3	8M9	Medium	MC	Numbers
1	2	1	70	1	4	8M9	Easy	SR	Numbers
1	2	1	107	1	5	8M9	Hard	MC	Data and Stats
1	3	0	52	0	6	8M9	Hard	MC	Geometry
2	2	2	132	1	7	8M9	Easy	SCR	Algebra
1	5	1	275	1	8	8M9	Hard	MC	Measurement
0	5	3	232	3	9	8M9	Hard	SCR	Geometry
1	2	1	51	1	10	8M9	Hard	MC	Numbers
0	4	3	135	7	11	8M9	Medium	MC	Algebra
2	3	0	92	0	12	8M9	Hard	SCR	Numbers
0	4	2	223	2	13	8M9	Hard	SCR	Algebra
0	5	2	155	1	14	8M9	Hard	MC	Data and Stats
0	0	0	0	0	15	8M9	Medium	MC	Algebra
0	0	0	0	0	16	8M9	Hard	MC	Numbers
0	0	0	0	0	17	8M9	Medium	MC	Algebra
0	0	0	0	0	18	8M9	Hard	MC	xxx
0	0	0	0	0	19	8M9	Easy	MC	Algebra
1	4	3	94	5	1	8M3	Easy	MC	Numbers
1	2	0	46	0	2	8M3	Easy	SR	Data and Stats
1	2	0	37	0	3	8M3	Medium	SCR	Numbers
2	2	0	34	0	4	8M3	Easy	SR	Algebra
2	2	0	50	0	5	8M3	Easy	SR	Measurement
0	3	0	86	0	6	8M3	Medium	SR	Algebra
0	4	0	224	0	7	8M3	Hard	SCR	Measurement
1	4	0	78	0	8	8M3	Hard	MC	Geometry
1	2	0	47	0	9	8M3	Easy	SR	Measurement
1	4	0	135	0	10	8M3	Hard	SR	Geometry
0	4	3	341	2	11	8M3	Hard	SCR	Algebra
0	2	0	44	0	12	8M3	Hard	SR	Geometry
0	3	3	217	18	13	8M3	Hard	ECR	Data and Stats
1	5	3	182	5	14	8M3	Hard	SR	Data and Stats
0	4	0	186	0	15	8M3	Hard	SCR	Geometry

Note: In the item type column (Type), MC = multiple choice, SR = composite selected response, SCR = short constructed response, and ECR = extended constructed response.

Table A.10. A representative instance from Profile 9: High performing and regulated time and tool uses.

	Time_M3 (min)	Time_M9 (min)	Total Score	Total Tool Use
Rep	24.47	29.15	29.00	22.00
Clust_mean	23.14	24.73	30.29	23.17
clust_std	4.83	4.28	3.21	10.74

Score	Time	Tool	Raw_Time	Raw_Tool	Item#	Block	Difficulty	Type	Content_short
1	3	3	31	3	1	8M9	Medium	MC	Measurement
1	4	2	141	1	2	8M9	Easy	MC	Geometry
1	2	0	34	0	3	8M9	Medium	MC	Numbers
1	2	1	98	1	4	8M9	Easy	SR	Numbers
1	3	1	118	1	5	8M9	Hard	MC	Data and Stats
0	2	0	46	0	6	8M9	Hard	MC	Geometry
0	2	0	105	0	7	8M9	Easy	SCR	Algebra
0	3	1	135	1	8	8M9	Hard	MC	Measurement
1	4	3	160	3	9	8M9	Hard	SCR	Geometry
1	4	3	159	3	10	8M9	Hard	MC	Numbers
1	2	3	49	3	11	8M9	Medium	MC	Algebra
2	3	2	105	1	12	8M9	Hard	SCR	Numbers
1	4	2	148	1	13	8M9	Hard	SCR	Algebra
0	2	2	60	1	14	8M9	Hard	MC	Data and Stats
1	2	0	35	0	15	8M9	Medium	MC	Algebra
1	3	0	23	0	16	8M9	Hard	MC	Numbers
0	5	3	102	3	17	8M9	Medium	MC	Algebra
1	4	0	118	0	18	8M9	Hard	MC	Data and Stats
1	3	0	57	0	19	8M9	Easy	MC	Algebra
1	2	0	14	0	1	8M3	Easy	MC	Numbers
1	2	0	47	0	2	8M3	Easy	SR	Data and Stats
1	5	0	180	0	3	8M3	Medium	SCR	Numbers
2	2	0	51	0	4	8M3	Easy	SR	Algebra
2	3	0	113	0	5	8M3	Easy	SR	Measurement
1	4	0	173	0	6	8M3	Medium	SR	Algebra
0	2	0	53	0	7	8M3	Hard	SCR	Measurement
1	2	0	36	0	8	8M3	Hard	MC	Geometry
1	4	0	155	0	9	8M3	Easy	SR	Measurement
1	3	0	78	0	10	8M3	Hard	SR	Geometry
0	2	0	159	0	11	8M3	Hard	SCR	Algebra
0	2	0	38	0	12	8M3	Hard	SR	Geometry
1	3	0	206	0	13	8M3	Hard	ECR	Data and Stats
2	3	0	93	0	14	8M3	Hard	SR	Data and Stats
0	2	0	46	0	15	8M3	Hard	SCR	Geometry

Note: In the item type column (Type), MC = multiple choice, SR = composite selected response, SCR = short constructed response, and ECR = extended constructed response.

Table A.11. A representative instance from Profile 10: Highest-performing group and nearly full time, with low tool uses.

	Time_M3 (min)	**Time_M9 (min)**	**Total Score**	**Total Tool Use**
Rep	30.28	29.70	40.00	6.00
Clust_mean	23.91	23.21	38.12	12.04
clust_std	4.82	5.71	2.72	3.86

Score	Time	Tool	Raw_Time	Raw_Tool	Seq#	Block	Difficulty	Type	Content_short
1	4	0	57	0	1	8M9	Medium	MC	Measurement
1	2	0	39	0	2	8M9	Easy	MC	Geometry
1	2	1	49	1	3	8M9	Medium	MC	Numbers
1	2	1	103	1	4	8M9	Easy	SR	Numbers
1	2	1	106	1	5	8M9	Hard	MC	Data and Stats
0	4	0	76	0	6	8M9	Hard	MC	Geometry
2	3	0	177	0	7	8M9	Easy	SCR	Algebra
0	5	1	408	1	8	8M9	Hard	MC	Measurement
1	2	0	53	0	9	8M9	Hard	SCR	Geometry
1	2	0	34	0	10	8M9	Hard	MC	Numbers
1	3	0	77	0	11	8M9	Medium	MC	Algebra
2	3	0	122	0	12	8M9	Hard	SCR	Numbers
1	3	2	117	1	13	8M9	Hard	SCR	Algebra
1	2	2	57	1	14	8M9	Hard	MC	Data and Stats
1	2	0	32	0	15	8M9	Medium	MC	Algebra
1	4	0	40	0	16	8M9	Hard	MC	Numbers
1	5	0	134	0	17	8M9	Medium	MC	Algebra
1	3	0	45	0	18	8M9	Hard	MC	Data and Stats
1	2	0	21	0	19	8M9	Easy	MC	Algebra
1	2	0	15	0	1	8M3	Easy	MC	Numbers
1	2	0	54	0	2	8M3	Easy	SR	Data and Stats
1	2	0	51	0	3	8M3	Medium	SCR	Numbers
0	3	0	65	0	4	8M3	Easy	SR	Algebra
2	2	0	73	0	5	8M3	Easy	SR	Measurement
1	2	0	72	0	6	8M3	Medium	SR	Algebra
1	3	0	85	0	7	8M3	Hard	SCR	Measurement
0	2	0	41	0	8	8M3	Hard	MC	Geometry
1	3	0	63	0	9	8M3	Easy	SR	Measurement
2	4	0	101	0	10	8M3	Hard	SR	Geometry
2	4	0	324	0	11	8M3	Hard	SCR	Algebra
2	4	0	111	0	12	8M3	Hard	SR	Geometry
4	5	0	513	0	13	8M3	Hard	ECR	Data and Stats
2	5	0	189	0	14	8M3	Hard	SR	Data and Stats
1	2	0	47	0	15	8M3	Hard	SCR	Geometry

Note: In the item type column (Type), MC = multiple choice, SR = composite selected response, SCR = short constructed response, and ECR = extended constructed response.