Understanding Students' Problem-Solving Processes via Action Sequence Analyses

Ruhan Circi
American Institutes for Research

Manqian Liao
Duolingo

Chad Scott
Deloitte

Juanita Hicks
American Institutes for Research

Research in this paper was developed and conducted during the 2019 NAEP doctoral internship program administered by AIR and funded by NCES under Contract No. ED-IES-12-D-0002/0004. The views, thoughts, and opinions expressed in the paper belong solely to the authors and do not reflect NCES position or endorsement.

ABSTRACT

The transition of the National Assessment of Educational Progress (NAEP) to digitally based assessments (DBAs) allowed for the collection of data that can provide insights into students' problem-solving processes. When students interact with a NAEP DBA item, their recorded timestamped events in the process data form sequences. We refer to action sequences as the series of clicks or other actions a student makes within an item. Using data from one released block of the NAEP 2017 mathematics assessment for grade 4, this study aims to provide an understanding of the relationships among action sequence characteristics, item characteristics and student performance.

First, we extract individual actions sequences across items. Second, we categorize each individual action into one of four activities: Browsing, Passive investigation, Active investigation, or Decision. This categorization enables us to investigate sequence patterns within and across different items. Sequence characteristics are summarized from two perspectives: a) the time spent on each activity is calculated for each student across items and b) the within-sequence entropy (Shannon, 1948) and turbulence (Elzinga, 2006) of the sequences are calculated to quantify students' action mobility.

We found that the time students spend on "Decision" and "Passive investigation" activities can be used to predict student performance.

Keywords

NAEP, Process data, Digitally based assessments, sequence mining, action sequences.

1. BACKGROUND

In 2017 the National Assessment of Educational Progress (NAEP) transitioned from paper-based assessments (PBAs) to digitally based assessments (DBAs). DBAs allow us to capture student interactions with the test screen that are recorded as timestamped events. These records form data known as process data.

It has become commonplace to include response time (RT) in addition to responses in the psychometric models to account for speed and accuracy (e.g., Goldhammer, 2015), and to examine the relationship between response time and item- and person-level factors (e.g., Masters, Schnipke, & Connor, 2005). Response time is used to examine psychometric quality of items and students' test-taking behaviors and it is concluded to be promising for various assessment elements. Yet, the process data contains richer information such as actions that students use during their problem-solving processes and the allocation of the time students spend on particular activities within a single response time remained unexplored.

When students interact with a NAEP DBA item, their recorded timestamped events in the process data form sequences. These sequences contain information about the order, mobility, and duration of the tasks students take throughout the problem-solving process and may shed light on the processes underlying the students' test-taking behaviors. In this study, we divide response times into subcategories using the action definitions to provide a more meaningful understanding of student test taking behavior and examine the differences across item types and student performance.

1.1 Literature

Process data is most commonly used to calculate response time (RT), defined as the time an examinee takes to complete an item or assessment. Due to the association of RT with psychological and cognitive processes (e.g., Huff & Sireci, 2001), RT is often used to make decisions such as setting assessment time limits (e.g., van der Linden, 2011) and capturing aberrant test-taking behaviors (e.g., Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014).

However, RT alone may not provide sufficient information to draw inferences about the processes underlying students' test taking behaviors (Lee & Haberman, 2016). In fact, RT could consist of the time for various components in the problem-solving process such as preparation (e.g., forming a response plan) and writing down/typing the response. The decomposition of RT can differ depending on item types (e.g., Li, Banerjee, & Zumbo, 2017). Thus, to ensure the validity of inferences drawn from RTs, it is necessary to understand what students actually do throughout the RT.

2. CURRENT STUDY

In the assessment setting, RT could consist of the times for various components in the problem-solving process such as preparation (e.g., forming a response plan) and writing down/typing the response. The decomposition of RT can be different for different item types (e.g., Li, Banerjee, & Zumbo,

2017). Since the NAEP mathematics assessment consists of items with a mix of item types (e.g., multiple choice, constructed response), using the decomposition of RT for different tasks (e.g., investigation, decision) rather than total RT could be helpful when different decisions (e.g., setting assessment time limits, capturing aberrant test-taking behaviors) are to be made based on the time.

A more fine-grained understanding of the relationships among RTs and students' problem-solving behaviors can be gained by analyzing students' action sequences, which can further improve the usefulness of RT in psychometric research (e.g., determining non-response categories such as omit and not reach).

The goals of this study are: a) identify and describe the action sequences of students in a meaningful way, b) examine mobility across actions, c) differentiate profiles of action sequences, and d) explore students' performance in connection to sequence clusters.

Steps taken for current project can be presented as follows: First, individual actions are extracted. Second, students' response processes are represented as sequences consisting of four tasks, i.e., Browsing, Passive investigation, Active investigation, and Decision (See definitions in Table 1). Since the variation across time for individual actions can be very large, we decided to use a set cut off point (2 seconds) for defining each action. In the end, we recoded the sequence of student actions in these groups for further analyses (See Figure 1 for an example). Then, the characteristics of the sequences are summarized from two perspectives: a) the time spent on each task is calculated for each student, which allows the decomposition of the RT, and b) the within-sequence entropy (Shannon, 1948) and turbulence (Elzinga, 2006) of the sequences are calculated to quantify students' action mobility.

Table 1. Definition and Example Action of Each Behavior Category

Behavior category	Definition	Example Action
Browsing	Examinees browse the content of an item by executing scroll on the screen	Horizontal scrolling, vertical scrolling
Passive investigation	Examinees get support from assistive tool for their problem-solving process without interacting item	Change theme (change the color of background)
Active investigation	Examinees interact with item as a part of their problem-solving process	Draw with scratchwork, highlight
Decision	Examinees make responses to an item	Click choice, text enter

In addition to summarizing sequence characteristics in a descriptive manner, this study examines the relationships among the sequence characteristics, item characteristics and students' item responses. Specifically, to examine the relationship between sequence characteristics and item characteristics, the RT decomposition and students' action mobility are compared across different items. Furthermore, representative sequence(s) are identified for each item with the use of a sequence dissimilarity measure and a clustering algorithm. The representative sequence(s) can inform the typical response process of an item.

Finally, to examine the relationship between sequence characteristics and student performance, sequence characteristics, such as the time duration of each task, within-sequence entropy and turbulence, are used as features to predict students' item scores. The results could inform which feature(s) of the sequences best contribute to correct/incorrect item responses or the presence/absence of the responses. Moreover, the score distributions are compared across sequence clusters.

In sum, this study, by decomposing RT and examining the relationships among the sequence characteristics, item characteristics and student performance, aims to inform more meaningful ways of calculating RT (e.g., different ways of RT calculation for different items) and the validity of score categories such as "omit" and "not reach". For instance, if the sequences of students who were scored as "not reach" were found to contain some actions that are related to making responses (i.e., the "Decision" actions), the scores of these students may be considered as "omit" as opposed to "not reach".

2.1 Research Questions

Specifically, the following research questions are examined in the current study:

RQ1. What actions do students take and what are the characteristics of the action sequences (mobility, time distribution) throughout the RTs of the NEAP math items?

RQ2. How do students' action sequences differ across different item types (e.g., multiple-choice item, constructed-response item)?

RQ3. Which action sequence characteristic(s) best predict the item scores?

3. DATA

We used data from one of the released blocks from NAEP 2017 Grade 4 Mathematics assessment. One of the released blocks includes 29,100 4th graders in both public and private schools and consists of 14 cognitive items. The sample was collected using the conventional NAEP sampling procedures, i.e., a two-stage stratified random sampling design with schools selected in the first stage and students in the second stage. In the data cleaning procedure, students with accommodation or interruptions were excluded. Comparisons of the demographic composition of the two samples, full sample and analytical sample, are presented in Table 2.

Table 2. Summary Statistics for Full and Analytical Sample: Student Demographic Characteristics

	Weighted		Unweighted	
	Analytical	**Full**	**Analytical**	**Full**
Observations	649,500	780,500	24,100	29,100
Gender	Percentages			
Female	50	49	50	49
Race/Ethnicity	Percentages			
White	51	49	52	50
Black	14	15	17	18
Hispanic	24	26	20	22
Asian	6	5	4	4
American Indian	1	1	2	2
Other	4	4	5	5
National School Lunch Program*	Percentages			
Eligible	48	50	51	54
Not Eligible	46	44	45	43

* No Information categories are not presented.
Note: Because all extended time accommodation students (that are excluded from analyses) are either with limited English

proficiency or in individualized education program, the results for these variables are not included. Detail may not sum to totals because of rounding.

A small non-significant difference in the proportion of White (50.5 % in analytical, and 48.9% in full sample) and Hispanic students (24.4% in analytical and 25.9% in full sample) are observed. A significant difference in term of NSLP non-eligible category is found (46% vs. 43.8%).

4. ANALYSIS

To construct sequences and decompose RT from the process data, we followed two steps (See Figure 1 for a demonstration of the procedure): a) Recoding the actions into four task categories (i.e., Browsing, Passive investigation, Active investigation, Decision; See definitions in Table 1); and b) Calculating the time duration of each task. Thus, students' item response processes were represented as sequences whose lengths are proportional to the time durations. Since the variation of time students spend on an item can be large (i.e., range from 0.01 second to 30 minutes), using a small-time unit (e.g., 0.01 second) could result in extremely long sequences that exceed the computer memory capacity. Therefore, we decided to use 2 seconds as the time unit while constructing the sequences. Only actions in students' initial item visit (i.e., actions between the first pair of "Enter Item" and "Exit Item" actions) were included in the sequence. Students whose initial item visit lasts longer than 8 minutes (480 seconds) were excluded from the analyses to avoid extremely long sequences. For all the items in the MA block, the percentages of students with initial item visit longer than 8 minutes are lower than 1%.

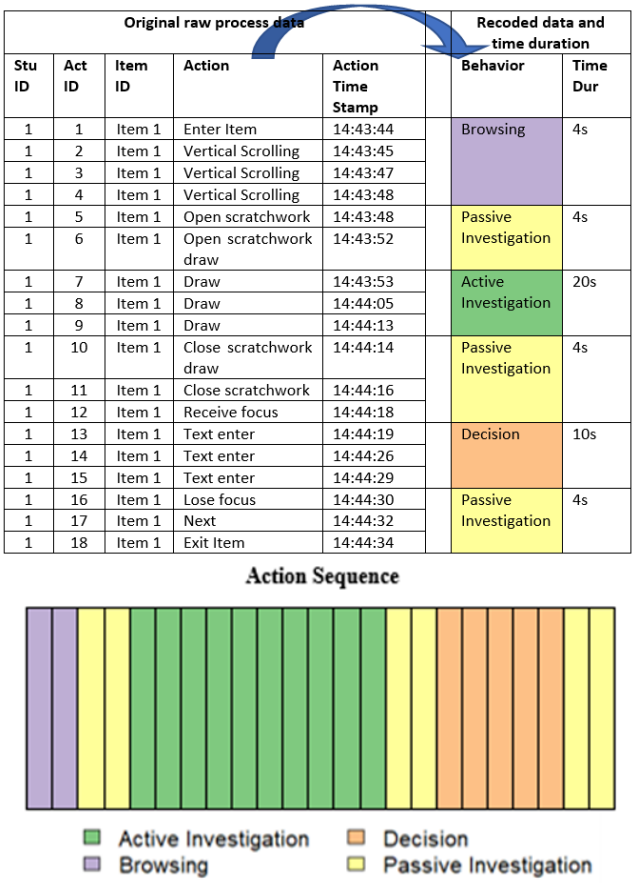

Figure 1. The procedure of turning raw process data into an action sequence.

The mean time spent on each action as well as the action mobility were summarized as the sequence characteristics. The number of task transitions, Shannon entropy (Shannon, 1948) and turbulence (Elzinga, 2006) measures were used to quantify the action mobility.

To examine how students' action sequences differ across different item types (e.g., multiple-choice item, constructed-response item), the characteristics of sequences were summarized and compared across different items. To identify the typical response process for an item, the hierarchical agglomerative clustering algorithm was applied to all the students' sequences based on the optimal matching edit distance (Levenshtein, 1966) matrix. The medoids of the clusters (i.e., the sequence that is the nearest to the virtual center of the cluster) were treated as the representative sequences that represent the typical response processes for an item. As no study to our knowledge has been done to determine the optimal number of clusters when the clustering is based on the edit distance matrix. Ward's algorithm was used to form clusters by maximizing within cluster homogeneity. We chose the number of clusters by visually inspecting the dendrogram and assessing the interpretability of the clusters. Specifically, for each item, we examined the cluster medoids when the number of clusters ranged from 2 to 4 and chose the number of clusters that resulted in interpretable clusters from practical perspectives. All sequence analyses were performed using the TraMineR R package.

To examine the relationship between the sequence characteristics and student performance, the sequence characteristics were used as features to predict the item scores using the regression tree (Breiman, 2017). In addition, the score distributions were compared across the sequence clusters identified based on the hierarchical clustering algorithm and edit distance.

5. RESULTS

For the purposes of this paper, we present the results for two selected items[1] listed in Table 3. The items are different item types (Item A is multiple-choice item while Item B is constructed response item) and are close in the presentation order. Thus, the two items were chosen to demonstrate the difference in sequence characteristics between items of different types (with minimal confounding of the presentation order).

Table 3. Characteristics of the Two Example Items

Item Characteristics	Item Label	
	Item A	Item B
Item type	Multiple-Choice	Fill In Blank
Presentation order	2	4
Item difficulty parameter	-0.17	0.29
Item content description	Compare heights of objects in a figure	Divide 3-digit whole number by 1-digit whole number

5.1 Response Time Decomposition

The average time students spent on each recoded behavior actions, i.e., browsing, passive investigation, active investigation, and

[1] https://nces.ed.gov/NationsReportCard/nqt/Search

decision are shown in Figure 2. For Item A, the "decision" task had the highest average time among the four tasks; however, for Item B, the "passive investigation" task had the highest time. On average, students spent 10 seconds browsing Item A by executing scroll on the screen, while students hardly spent any time browsing Item B. Such difference in the browsing time could be associated with the content of the items: Item A needs to be solved by inspecting and comparing the heights of the trees which may result in browsing actions, while Item B is a straightforward computational item which may not require much browsing.

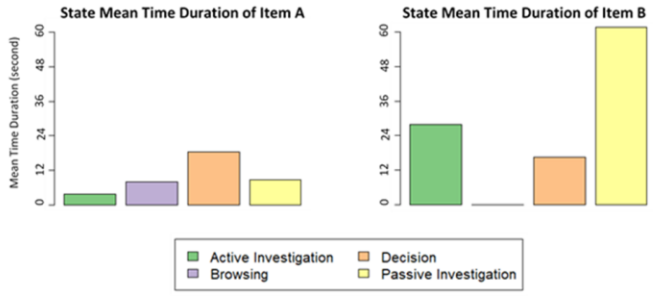

Figure 2. Average time students spent on recoded actions when interacting with two selected items.

5.2 Sequence Characteristics

Figure 3 presents the state distributions at each time unit for the two selected items. Each unit of the x-axis represents 2 seconds. For instance, in the first 2 seconds, students who conducted "passive investigation" make up the largest proportion in both items. When responding to Item A, more than 10% of the students were browsing the item in the first 2 seconds; when interacting with Item B, nearly no students browsed the item in this time unit.

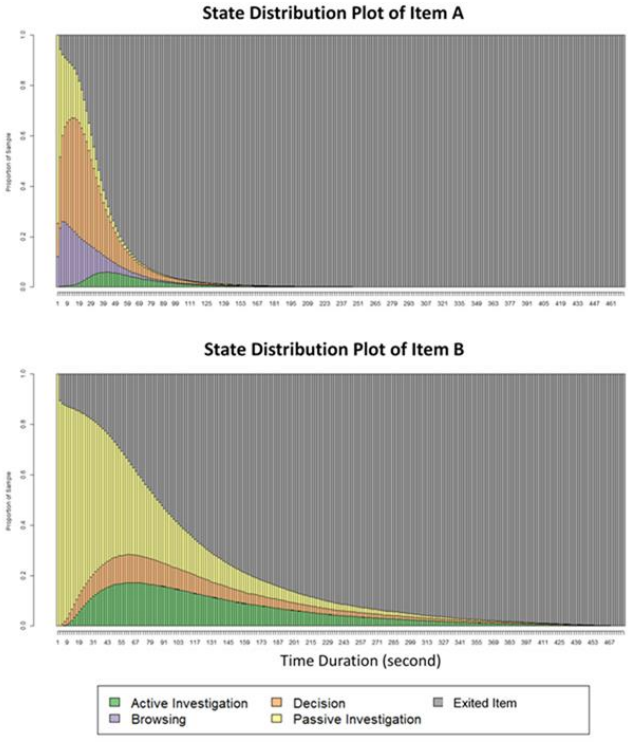

Figure 3. State distribution plot of the two selected items.

Table 4 lists the summary statistics of three mobility measures, i.e., the number of task transitions, within-sequence entropy, and turbulence. Task transition refers to switching among the four tasks (i.e., browsing, passive investigation, active investigation, and decision) in the sequence. The average task transitions for item A and B are 2.28 and 2.13, respectively. As for the within-sequence entropy and turbulence measures, higher values indicate larger mobility. On average, item A is found to have higher within-sequence entropy and turbulence.

Table 4. Mobility Measures of the Two Selected Items

Mobility Measure		Item A	Item B
Number of task transitions	Min	1	1
	Median	2	2
	Mean	2.28	2.13
	Max	4	4
Within-Sequence Entropy	Min	0	0
	Median	0.38	0.29
	Mean	0.37	0.30
	Max	1	0.92
Turbulence	Min	1	1
	Median	3.18	2.76
	Mean	3.36	3.28
	Max	11.24	14.43

5.3 Typical Response Process

Figure 4 shows the representative sequences for Item A and Item B. A representative sequence refers to the sequence with the smallest sum of edit distance to the rest of sequences; the representative sequence is considered to be representative of the typical response process of an item. As the sequence length is proportional to the time duration, the overall time duration of the typical response process is shorter for Item A than Item B. We observe that, in the typical response process;

- for Item A, the student conducts passive investigation, browses the item, and makes response decisions, sequentially.
- for Item B, the student conducts passive and active investigations and makes response decisions.

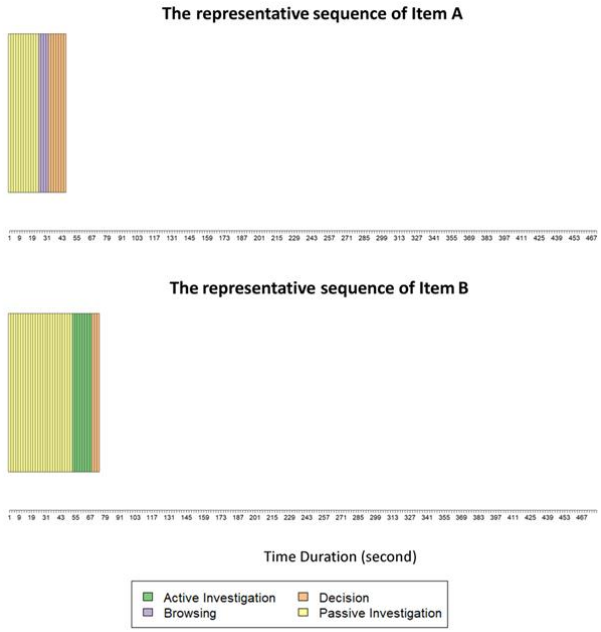

Figure 4. Representative sequences of the two selected items.

While identifying a single typical response process for an item is desirable for the purpose of interpretation, a single sequence may not be enough to represent all the sequences. It is possible that there are multiple response process archetypes for an item. Thus, we conducted hierarchical agglomerative clustering based on the edit distance matrix. In the clustering process, each unit is a student. After examining the dendrogram and the interpretability of the clusters, we chose to retain three clusters (labeled as Type 1, Type 2, and Type 3). The weighted cluster sizes and students' demographic characteristics by sequence clusters found in Item B are presented in Table 5.

Table 5. Student demographic characteristics by Sequence Clusters in Item B

	Weighted Percentages		
	Type 1	**Type 2**	**Type 3**
Observations	307,100	172,500	157,400
Gender	Percentages		
Female	45	52	57
Race/Ethnicity	Percentages		
White	49	53	51
Black	15	13	13
Hispanic	24	23	26
Asian	6	6	4
American Indian	1	1	1
Other	4	5	4
National School Lunch Program*	Percentages		
Eligible	48	45	49
Not Eligible	45	49	45

Detail may not sum to totals because of rounding.

Figure 5 displays the representative sequences of the three clusters in Item B, which represent three archetypes of response processes in this item. The representative sequences of Type 1 and Type 3 only consist of "passive investigation" and "decision". The time duration of "passive investigation" is longer for the representative sequence in Type 3 than Type 1. The representative sequence of Type 2 consists of "active investigation" in addition to "passive investigation" and "decision".

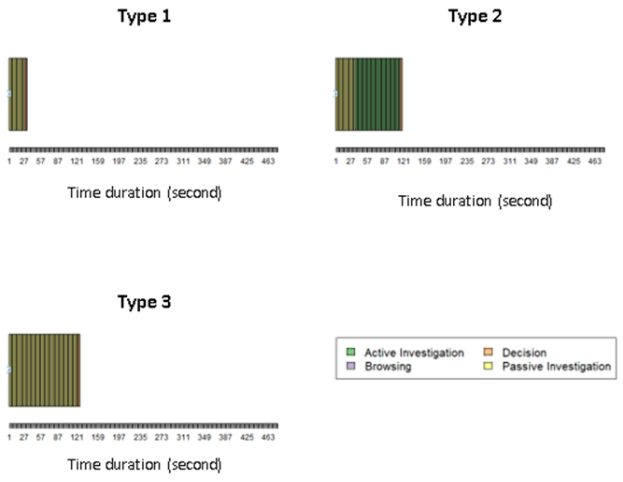

Figure 5. Representative sequences of the three sequence clusters found in Item B.

5.4 Relationship Between the Sequence Characteristics and Student Performance

Figure 6 shows the regression tree learned from the process data of Item B. Time durations of browsing, passive investigation, active investigation and decision, number of task transitions, within-sequence entropy and turbulence are used to predict item scores. Item B has five score categories, i.e., incorrect, correct, off task, omitted, and not reached. Each box in Figure 6 is called a "node" and the five decimals in each box are the predicted proportions of students having the five score categories in that node. The name (and color) of the node is determined by the score category that has the highest proportion among the five categories. For example, as the first split was performed with the decision time, for students with decision time longer than 14 seconds (25% of the students in the sample have decision time longer than 14 seconds), the predicted proportions of getting "incorrect" and "correct" scores are 0.70 and 0.28, respectively. In addition, all the splits in this regression tree are performed with either decision or passive investigation time durations.

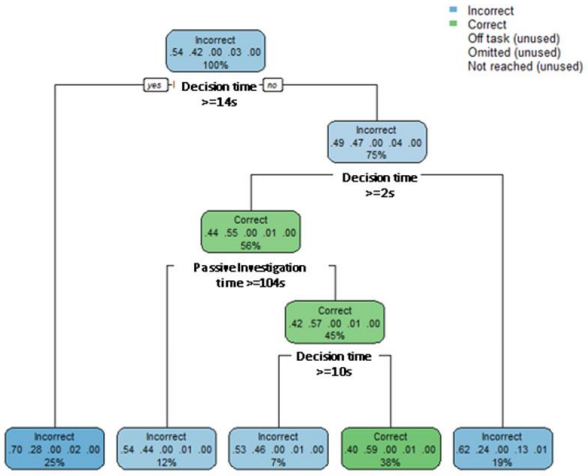

Figure 6. Regression tree learned from the process data of Item B.

5.5 Relationship between the Sequence Cluster and Student Performance

Table 6 lists the score distribution of scores within each sequence cluster found in Item B. Item B is a fill-in-blank item, which is the fourth item in the block with a difficulty level of 0.29. In all three clusters, the proportion of students getting "incorrect" score was the highest among the five score categories. The similarity in the score distributions across sequence clusters implies that no clear pattern on the performance difference has been found among students with different response process archetypes.

Table 6. Score Distribution of Each Sequence Cluster in Item B.

		Percentages (%)				
Cluster	**Cluster Size**	**Correct**	**Incorrect**	**Omitted**	**Not reached**	**Off task**
Type 1	11,500	42.2	54.0	3.4	0.2	0.2
Type 2	6,100	42.9	53.5	3.3	0.2	0.1
Type 3	5,900	42.5	53.7	3.5	0.2	0.2
Note. Percentages of each row add up to 100%.						

6. DISCUSSION

6.1 Summary

In summary, this study provided insights into the decomposition of RT by constructing action sequences from students' process data. In particular, the action sequences contained information of the time duration, order and mobility of the tasks students executed to solve the NAEP mathematics items. By presenting the sequences of two selected NAEP released items as examples, this paper demonstrated the differences in RT decomposition and typical response processes between items of different types (i.e., a multiple-choice item vs a fill-in-blank item). This methodology and set of results suggest that examining action sequences and RT decomposition can be a useful way to mine process data and uncover educational processes. Also, action sequence mining can be useful to analyze high variance data such as process data.

Response process archetypes were found by conducting a hierarchical clustering algorithm using the edit distance matrix of students' action sequences. As for the relationship between student performance and sequence characteristics, the time students spent on "Decision" and "Passive investigation" were incorporated in the learned regression tree of the example fill-in-blank item, meaning that these components of RT can be used to predict the scores of this item. Further, among the 10,000 students who correctly responded to Item B, 48.6% had their action sequences clustered into Type 1, 26.3% into Type 2 and 25.2% into Type 3, which implied that students who responded to the item correctly may have different response processes.

6.2 Limitations and Future Research

As an initial exploration of action sequences in the NAEP mathematics items, this study has limitations and opens up opportunities for future research. First, the actions were categorized into four tasks (browsing, passive investigation, active investigation and decision) in this study. However, this may not be the only way to categorize the actions. For instance, in a multiple-choice item, the actions could be recoded based on students' selected options. Thus, sequences that reflect students' trajectory of answer changes can be constructed.

Second, the number of clusters was determined only based on the dendrogram and the interpretability of the clusters in this study. To better justify the choice of the number of clusters, future studies could develop quantitative measures to determine the optimal number of clusters based on the edit distance matrix.

Finally, this study only included a limited number of sequence characteristics as features to learn the regression tree. Other features such as the frequencies of subsequences (e.g., the frequency of a student switching from passive investigation to active investigation and then to decision), together with feature selection algorithms, could be incorporated in future studies.

7. REFERENCES

[1] Breiman, L. (2017). *Classification and regression trees*. Routledge.

[2] Elzinga, C. H. (2006). Turbulence in categorical time series. *Mathematical Population Studies*.

[3] Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives, 13*(3–4), 133–164.

[4] Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16–25.

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

[6] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

[7] Li, Z., Banerjee, J., & Zumbo, B. D. (2017). Response time data as validity evidence: Has it lived up to its promise and, if not, what would it take to do so. In B.D. Zumbo & A.M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 159–177). Springer.

[8] Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*(6), 426–451.

[9] Masters, J., Schnipke, D. L., & Connor, C. (2005). Comparing item response times and difficulty for calculation items. *In annual meeting of the American Educational Research Association, Montreal, Canada*.

[10] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379–423.

[11] van der Linden, W. J. (2011). Setting time limits on tests. *Applied Psychological Measurement, 35*(3), 183–199.