Introduction to Process Data

Ruhan Circi, Ph.D.
Center for Process Data, American Institutes for Research (AIR)

Transcript

Transcription is provided in order to facilitate communication accessibility and may not be a totally verbatim record of the proceedings.

(Slide 1)

Hello everyone. Welcome to the NCSER "Introduction to the NAEP Process Data" webinar. I am Ruhan Circi, your presenter for today. And I am the Director of the Center for Process Data. This webinar is provided as a joint task between NCSER, NCES and AIR to provide an introduction to the NAEP Process Data, and how it enables research in ways not possible before.

This presentation will introduce NAEP Process Data. We hope by the end of this webinar you will have a clearer idea about the basics of NAEP process data and information about specific research examples. You are welcome to submit your questions up until July 30th at IES.Webinar.Questions@ed.gov.

Updates will be posted to the frequently asked questions document on the webinar website every few weeks. I also would like to mention that we will simplify some of the concepts in our technical language as much as possible during this webinar.

(Slide 2)

We are going to talk about three big topics today.

First, we will provide an overview of process data, which includes how assessment has changed in the age of the digital era, how NAEP data is created, what exactly it is, and how it is embedded in NAEP mathematics assessments.

Next, we will have a deep dive into the data, getting a glimpse of its unique features, the actions that are captured, as well as examples of captured and user-generated variables.

Finally, we will highlight examples of research projects using process data.

(Slide 3)

Now, advances in digital technologies allow for a wider scale and more rich data gathering, processing and analysis. All sides of education have been impacted by this rapid evolution. There have been innovative improvements in nearly all areas of the assessment cycle, from item development to the type of feedback provided based on the assessment results.

Currently, many areas of the assessment cycle are enhanced and semi-automatic, where data cuts across most lines of the topics. Looking towards the future, assessments could become fully data-driven and heavily used to infer the behavioral and cognitive processes displayed by examinees. So, the future requires a different perspective, let's see how process data can inform and contribute to it.

(Slide 4)

Let's start with an overview of NAEP Process Data.

(Slide 5)

NAEP transitioned to the digitally based assessments in reading and mathematics in 2017. The NAEP Process Data Snapshot is a quick and simple way to understand what process data is at its most basic level.

First, students received a digitally based assessments on tablets. Then, students' interactions with the assessment system and tablet are logged in the background. Some example actions are “zooming in and scrolling vertically”. For each action, accompanying information is logged, such as the timestamp and item type, for example, Multiple Choice Single Select.

Thus, there is a traceable receipt for each student's assessment experience. This data is then formatted for further purposes into an easier to read data file, where an example is included in the bottom right corner.

Next, let's see how process data is embedded in the NAEP assessment design.

(Slide 6)

Before we get into the details of process data, it is important to understand the parts of the NAEP assessment, where data is being collected. The current design of NAEP assessments has six general sections: Adjustment; Tutorial; General Directions; Cognitive Section, which includes its own set of directions; Student Questionnaire, which also includes its own set of directions; and finally, the Thank You screen, which is the end of the assessment.

The blue boxes are the parts of the assessment we are most interested in when it comes to Process Data. In the tutorial section, students review tools and other helpful assessment features and information before the assessment begins, such as opening calculator or using highlight tool. In the cognitive section, students receive a subset of NAEP items for their particular grade and subject area, in this case, mathematics. These items are bundled in cognitive item blocks, and each block has a 30-minute time limit. In the case of students who received extended time accommodation, each cognitive block has a 90-minute time limit.

After the cognitive section, students received a student questionnaire, which has a variety of demographic and subject areas specific questions, such as, "How often do you use math in everyday life outside of school?"

Student actions or events are logged after a student has entered the system. It is worth mentioning that each section has separate actions recorded. For this webinar, we will focus on the process data within the cognitive section.

(Slide 7)

Let's look at what a digital assessment platform looks like.

This is an example of what students see when they go through the tutorial. There's a link to the most recent NAEP mathematics tutorial in the online resources document.

They are shown the system tools that can be used during the assessment. In the upper left corner, there are multiple system tools, including a help button which takes a student into the help page on system tools, a theme change button, zoom in and zoom out buttons, a text to speech button that activates Text to Speech/Read Aloud mode, a scratch work button that allows drawing and highlighting, as well as an equation editor and calculator button.

In the upper right corner, you can see the timer, the next and back arrows, and the navigation panel to answer items in any order that students choose to. Almost every action by the student or the system is stored in student's assessment log file. All of this data is thus labelled process data, which shows interactions between the student, the assessment platform, and the content presented.

In order to experience how these functionalities work, we suggest you watch the NAEP mathematics tutorial which is available online. Please note that the tutorial may have changed.

(Slide 8)

Now that you have a basic understanding of process data and where it comes from, we can look at the bigger picture which we call the Process Data Workflow.

This workflow, while still in development, provides a guidance on the major stages needed to fully understand and use process data properly. It is built upon data itself, yet there is more to it, such as Data Management, Security, Analytics, and Visualization. Each part of the framework is vital. However, for the remainder of this webinar, we will focus on Data and some example Analytics.

(Slide 9)

Let's deep dive into the data.

(Slide 10)

It is important to understand the structure of the data before starting to work with it. A file format is a standard way in which information or data is encoded for storage in a file.

For this webinar, we will explore the data in a comma separated value (csv) format. Please note that NAEP process data can be released in a different format, such as in plain text file format in the near future. It is also important to understand the data format. Currently, data is in the long format, where each row refers to an action or event that contains the type of the event, the time of the event, and the context in which the event occurred, such as item identifier (IDs).

Therefore, this data is different from conventional data sets where each student is represented by a unique row.

(Slide 11)

Let's take a look at what the variables in the process data refer to. Column A is the PseudoId, pseudo unique student identifier in this data set.

Column B is BlockCode, unique code for each section of the assessment, such as tutorial, directions, cognitive assessment and survey. Column C is AccessionNumber, or unique item identifiers, item IDs. For example, the cognitive section has unique item identifiers for directions for each item presented in the block, for Block review page, for help page, and for Timeleft message.

Column D is ItemTypeCode, that represents the type of the item. If it is a Multiple Choice Single Select item, we see the MCSS label. Column E is ObservableType, represents the observable events while a student is taking the assessment. For example, focus on the first green box.

The event indicates that the student entered an item. Column F is ExtendedInfo. Some events will capture extended information specific to that type of event. For example, focusing on the next green box, this student clicked on the TextToSpeech button, and we have extended information that TextToSpeech mode is on.

Column G is Timestamp. All observable events are captured along with a date timestamp. It is possible that data will be released with original timestamps, or with the cumulative time variable, as in the next column. Cumulative time variable is self-explanatory. It represents the cumulative time for each section in the assessment, that is, cumulative time resets to be zero when students finish the cognitive section and move to the survey section.

Please note that the availability of time related variables requires calculations by those performing the analysis, such as total time spent on each item.

(Slide 12)

Here's an example list of actions or events that are captured for the NAEP mathematics assessments. We have already seen some of these events on the previous slide within the ObserveableType column.

However, it is important to note that there are more events in the dataset. Some events that are captured are quite intuitive and self-explanatory, such as “Highlight”, “Increase Zoom”, and “Open Calculator”. Other events are not so intuitive. For example, “Calculator Buffer”, “DropChoice”, and “First Text Change”.

To help users know and understand what event data are captured and what they mean, a glossary is provided to accompany this webinar. Yet detailed documentation will be available when data is released in the future.

(Slide 13)

Let's take a deeper look at the data with an example student. On this slide, we see separate sections from one student's complete log. This student has 2285 rows across all sections, and 1678 rows in the cognitive section, which is presented in the middle of the slide.

We can see which tools this student used across different sections and items. We can also see how much total time the student spent on the cognitive section, in this case, 2204 seconds, it is almost 37 minutes. Given that each cognitive section has a 30-minute limit, can you think a reason why this student spent more than 30 minutes?

Yes, you're right, the student had an extended time accommodation.

(Slide 14)

Now, let's explore how data is aligned using student actions on one of the items. On the left side, there is a release item from 2017, Grade 8 mathematics assessment MC block on the NAEP questions tool. Item asked to identify a verbal description of a rotation of a figure in the xy-plane.

We will try to make sense of the observed actions or events that this student performed when interacting with this item. First, student enter this item. Then click on the TextToSpeech button to turn it on.

Then, click the TextToSpeech button to turn it off. Next, student click the TextToSpeech button to turn it on again. Student selected the items tab multiple times to be read aloud, indeed four times. Then, student selected option A to be read aloud, twice.

Next, student click the TextToSpeech button to turn it off. Then student select option A. Next, answer is cleared. Then student selected option E.

Student clicked Next button. And finally, student exited the item.

(Slide 15)

So far, we have looked into the process data and captured events. In addition to the captured actions, there are also assessment captured variables, such as student identifier (or pseudo ID in our data set), Block Code, Item identifier (or Accession Number), and Item type code.

The Observable Type variable stores the captured actions from slide 12. In NAEP data, the captured variables can be used to derive many more. Derived variables can be numeric (for example, cumulative time), nominal (for example, calculator use), or relational (for example answer change), among many others.

All it takes is a bit of creativity to explore what is needed for any given research question or questions. It is important to realize that this data can be used to derive new variables. Cumulative time is one of the derived variables you already see in the data set. Beyond this, it is possible to create new variables, such as number of visits by counting the how many times each student worked on each item, or whether the student used the calculator, lots of possibilities.

(Slide 16)

Let's see an example for derived variables.

Let's say that I am interested in extracting the frequency distribution of the number of total visits for a multiple choice item. Let's create a new variable and call it “Visit number”. With this new variable, we can examine descriptives and visualize the results.

We can see that maximum number of visits for this item was seven, but there were only a few students with that many visits.

(Slide 17)

Other information that will be available alongside processed data can include demographics, such as gender and parental education; accommodation types, such as extended time accommodation; as well as raw and scored data.

For example, option A in the raw score data for a Multiple Choice Single Select item will be scored as 1 if this option is correct. At the bottom of the slide, we see some more information about the student that we explored, her actions on slide 14. Please note that these are only a few of the example variables.

More student, school and teacher variables may be available for data release. Now, it is time to see some examples from current research projects.

(Slide 18)

While all this rich data can be used for reporting and for improving internal assessment operation processes, a big promise of it is it's used in innovative research.

(Slide 19)

The combination of all assessment events or actions, assessment variables, and assessment data leads to full-fledged research using process data to answer common and new research questions. Process data has the potential to gain insights into various assessment related areas, with perhaps its most vital use in quality assurance.

The totality of the process data uses, what we call the process data ecosystem. In the next section, we will highlight a few examples of research projects using process data.

(Slide 20)

One use of process data can be to identify student misconceptions, errors, or misunderstandings.

Currently, misconceptions are modelled using item and student performance data. For example, blank or skipped items may be classified as misunderstanding the item. With process data, item response time in conjunction with other information can identify if students spend time on items but ultimately choose not to answer them.

Information such as item type and item difficulty can further aid in modelling if an unanswered item is related to a misconception. Response change analysis can also shed light on possible student misunderstandings based on how answers are changed, or how different student groups change their answer.

Looking at the graph, the X-axis shows the item sequence for a block of items. And the Y-axis shows the percent of students who changed their answers at least once. You can see from the green line that across all items, accommodated students change their answers more compared to non-accommodated students.

(Slide 21)

Response change analysis can also enhance other common metrics used to address the misconceptions. Distractor analysis is commonly used to address the rate at which wrong answer alternatives are chosen. For example, 35% of students choose option C. Answer change analysis can tell you how many students changed their answer from one option to another, adding another layer to distractor analysis.

For students with accommodations, this could be especially useful to explore if there are items that are more of an inconvenience to accommodated students compared to non-accommodated students.

Looking at the graph, the X-axis shows the number of accommodated students, and the Y-axis shows the peer-wide response changes for item 11 from a NAEP's block. For example, you can see from the red boxes that there are students who changed their answers from correct or incorrect.

(Slide 22)

In another study, we explore the use of the extended time accommodation on performance. In paper and pencil tests, it is not easy to see how Extended Time Accommodation students use their extra time, if they do it at all.

With process data, it is easier to keep track of students with Extended Time Accommodation and how they use their extra time, if they use it. Explorations like this enhance the use, understanding, and addition of process data to research, especially student group specific research. Further analysis of extended time accommodation can also provide guidance on the optimum Extended Time Accommodation time.

Our results show that only one-third of Extended Time Accommodation students use their extra time. In the graph, the X-axis is the sum raw score for this block of items, and the Y-axis is the number of students. Explorations on extended time use in relation to the performance show that on average, students who use their extra time scored two points higher compared to students who didn't use their extra time.

(Slide 23)

Another insight that process data brings is in exploring the use of assessment features. Before currently designed assessments, features like text-to-speech were only reserved for certain students. Now, text-to-speech is a universal design feature that is available for all students to use.

Process data can show which students are using assessment features, and whether students who have traditionally had access to it (such as accommodated students), or not (such as non-accommodated students) use it. The graph shows about 38% of the non-accommodated students and 44% of the accommodated students used the text-to-speech feature at least once.

In all, what these example research projects show, there are many varied uses of process data above and beyond traditional analysis. This process data also provides insights into spatial student groups in a way that was not possible to do before.

(Slide 24)

Thank you for participating in this NCSER webinar on "Introduction to NAEP Process Data." In this webinar, we covered very basics of the NAEP process data and several research studies.

As a reminder, if you have any questions about this webinar, please send them to IES.Webinar.Questions@ed.gov address.

You are welcome to submit your questions up until July 30th. Updates will be posted to the Frequently Asked Questions document on the webinar website every few weeks.