Institute of Education Sciences
National Center for Education Statistics

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
TECHNICAL EXPERT PANEL REPORT

RELEASE OF PROCESS DATA TO RESEARCHERS

TABLE OF CONTENTS

NATIONAL INSTITUTE OF STATISTICAL SCIENCES

TECHNICAL EXPERT PANEL REPORT ON RELEASE OF PROCESS DATA TO RESEARCHERS

EXECUTIVE SUMMARY

Now that most National Center for Education Statistics (NCES) assessments and surveys are conducted using electronic modes, electronic data capture means that in addition to basic background information and final responses, data includes documentation of the process of responding. These process data comprise a time-stamped, click-by-click record of each student's progress through the assessment or survey. To date these detailed process data have not been made available; however NCES is now preparing to establish a mechanism for release of process data for research purposes. Therefore NCES asked the National Institute of Statistical Sciences (NISS) to convene panels of technical experts to advise them on how to most efficiently and safely release the data in a useful form. The first panel was charged with making recommendations about how to minimize disclosure risks from such a data release. The second panel was asked to consider what data would be most useful to researchers, and how it should be preprocessed to allow a broad range of users able to use the data efficiently. Specifically, this panel was asked for guidance on the creation of new variables that will be useful to researchers in the fields of education sciences, test development and psychometrics, behavioral psychology and related fields. The goal for NCES is to make the research process more efficient for these data users by preventing duplicative effort in creating these variables.

The panel met via teleconferences with an in-person meeting at NCES on 11-12 March, 2020. This summary discusses the second panel's recommendations.

Primary Recommendations

1. An internationally used data exchange standard should be adopted for raw process files.
   - Such standardization will allow the exchange among researchers of robust, reliable processing scripts.
   - Such standardization will have the added benefit of bringing in additional researchers interested in data mining, machine learning, etc.
2. For researchers who do not desire to work with raw process files, but who do want access to micro-data, we recommend preparation of two pre-processed files for each content area. One should be focused on items and one on examinees. For both item and examinee files, summary information comes in two forms: background information and summary statistics computed from the process data itself.
   - The examinee file should include:
     - Background variables:
       - A student ID that allows for matching to other examinee files
       - demographic variables, including SES variables
       - Disability information accommodation information

        - ELL status
        - Performance level and total points
        - School code
        - Complex sample design information necessary for population level estimation
        - Blocks the student took
    - Process summary data
        - Item path of the examinee's progress through the block, i.e., a block path file that maps the state of the user at each item, including time stamps, tools used, whether the question was answered/was correctly
            - For reading sections specifically, knowing the screen layout is critical - whether a student is viewing the prompt and the question (side by side or by page change), and the frequency of the change between views
        - # omitted, # not reached
        - Total time in block and %-ile of time for examinee
        - Examinee's answers to each item
        - Item visit count and total time in item
        - Item tool use for each type of tool
- The item file should include:
    - Background variables
        - Item ID that allows for matching to examinee file
        - Position in block and assessment
        - Item parameters
        - Subscale
        - # words, # images, #tables in item
        - Tool indicator (presence or absence of each tool available to user)
        - Correct response
        - Depth of knowledge measure
    - Summary variables
        - Summary statistics for time spent on item over examinees
        - Summary statistics for # of visits to item over examinees
        - % answer changed, omitted, and not reached over examinees
        - % tool use for each tool
        - Responses summary over examinees (e.g., % of each multiple choice selection)

3. A tool similar to or included within the NAEP data explorer should be developed for the process data for users who do not have interest/access to micro data.

PREFACE

Over the last few years, the National Center for Education Statistics (NCES) transitioned its large-scale assessments to electronic modes. As a result, they now capture data not only for the responses, but also for the behavior of the examinees, as they select and record their responses. These data include all clicks and text they enter, along with time-stamps of each. NCES would like to develop a strategy to make available as much of these data as possible to those who may find the data valuable for their educational research programs or practice. Therefore NCES asked the National Institute of Statistical Sciences (NISS) to convene panels of technical experts to advise them on how to most efficiently and safely release the data in a useful form. The first panel was charged with making recommendations about how to minimize disclosure risks from such a data release. The second panel was asked to consider what data would be most useful to researchers, and how it should be preprocessed to allow a broad range of users able to use the data efficiently. This report summarizes the second panel's recommendations.

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
TECHNICAL EXPERT PANEL REPORT

RELEASE OF PROCESS DATA TO RESEARCHERS

I. INTRODUCTION

The National Center for Education Statistics (NCES) now collects data for its large-scale assessment programs and associated surveys by computers or other electronic devices, such as tablets. As a result, they capture a new type of data now known in the assessment community as process data. These data include all the clicks and text that examinees enter on their assessment device, along with time-stamps of each. These data record detailed and previously unavailable information about the behavior of the examinees as they interact with the test items. NCES researchers and managers believe these data could be useful for providing insight into the cognitive processes of the examinees as they attempt to answer the questions. Research in this area might then inform better assessment design and methods of teaching and learning. Therefore, NCES is interested in releasing as much of the data as possible to the research community.

NCES requested the National Institute of Statistical Sciences (NISS) to convene a series of expert panels to consider what process data could be safely provided without incurring unacceptable risks, and to advise them on how to make the data available most efficiently, effectively, and securely. The first panel considered the question of how the released data can maintain privacy and confidentiality of examinees that NCES is obligated to provide. The recommendations from this panel were provided in their final report.

The second panel was charged with considering how to make the released process data valuable and efficient for researchers to use in their research. Specifically, the panel was asked to consider what variables should be created by NCES to reduce the effort researchers would need to preprocess the data before they could begin their research. They were also asked to suggest research questions that would be addressable with this new form of data. The full charge is available in the Appendix.

This report contains the panel's responses to this charge and recommendations for process data release. We focus most intensively on how to release process data from the 2017 Grade 8 Mathematics Assessment, as those data already have been prepared and are available for release. In fact, the first draft of these data are scheduled for release prior to completion of this report. Since use of process data in research is new to most education researchers, NCES is proactively planning for revisions based on user feedback. For example, this report's recommendations are designed for informing a second release of the 2017 Grade 8 Mathematics process data.

Our recommendations are presented in four additional sections. First, we make recommendations for release of the data to those researchers who have the training and the desire to analyze the files

containing the raw process data. We strongly recommend that NCES adopt a standard format for the process data products for this audience. The benefits of this approach and a discussion of options are included in Section 2. Section 3 focusses on specific recommendations for the release of data products for those users who prefer to receive data that have been pre-processed and are delivered in more familiar file formats. We discuss useful data summaries that NCES could provide to these users, especially for data from the Grade 8 Mathematics assessment, or other assessments with similar items. Section 4 provides recommendations for release of process data specifically for writing assessment. Though not imminent, process data from other content areas will eventually also be released, according to NCES. Finally, Section 5 discusses a range of research problems that we believe may be explored using the process data, either alone or together with the assessment cognitive data.

II. ADOPTING A DATA STANDARD

2.1 Overview and Purpose

There is a need for a fresh perspective for an assessment program's data governance, so that that it will efficiently support multiple functions. In this section, we discuss how to produce student data (process and outcome data) to be consistent with standards for interoperability and that will facilitate access to the same types of variables and data schemas. The goal is that this access will spur innovation in assessment technology by enabling researchers, institutions, and technology providers to efficiently collaborate and/or replicate each other's work using their own platforms and methodologies. The adoption of a data exchange standard will also more easily allow collaboration with researchers from other research communities, such as from the data mining and AI-based communities. This recommendation is based on the paper "The Argument for a Data Cube for Large Scale Psychometric Data" by von Davier et. al (2019).[1]

In recent years, work with educational testing data has changed due to improved technology, availability of process data collected in large data sets, and advances in data mining and machine learning. Consequently, data analysis moved from traditional psychometrics to computational psychometrics. In the computational psychometrics framework, psychometric theory is blended with large scale, data-driven knowledge discovery (von Davier 2017).

Many testing organizations like NCES have started to include the process data from the performance or activity-based tasks in the assessment, which led to new challenges around the data governance: data design, collection, alignment, and storage. Some of these challenges have similarities with those encountered and addressed in the field of learning analytics, in which multiple types of data are merged to provide a comprehensive picture of students' progress. For example, Bakharia et al. (2016), Cooper (2014), Rayon, et al. (2014) propose solutions for the interoperability of learning data coming from multiple sources. In recent years, the testing organizations started to work with logfiles and even before the data exchange standards for activities and events, such as the Caliper or xAPI standards, have been

[1] https://www.frontiersin.org/articles/10.3389/feduc.2019.00071/full

developed, researchers have worked on designing the data schema for this type of rich data (see Hao et al., 2014).

2.2 Data Exchange Standards

Data standards allow those interoperating in a data ecosystem to access and work with this complex, high-dimensional data (see for example, Cooper, 2014). Several data standards exist in the education space which allow testing and learning organizations to share information and build new knowledge. For example, users can combine test scores with the process data, items metadata, and demographics for each student in order to identify meaningful patterns that may lead to differentiated instructions or interventions to help students improve.

This makes it easier for those creating, transmitting, and receiving the data to avoid the need to create translations of the data from one system to the next. Data exchange standards allow for the alignment of databases (across various systems), and therefore, facilitate high connectivity of the data stored in the databases. Specifically, the data exchange standards impose a data schema (names and descriptions of the variables, units, format, etc.) that allow data from multiple sources to be accessed in a similar way.

Two data standards in the education space that address data exchange for process data are:

- IMS Global[2] Question & Test Interoperability Specification includes many standards. The most popular are the IMS Caliper and CASE.
    - IMS Caliper, which allows streaming in assessment item responses and processes data that indicate dichotomous outcomes, processes, as well as grade/scoring.
    - IMS Global Competencies and Academic Standards Exchange (CASE), which allows importing and exporting machine readable, hierarchical expressions of standards knowledge, skills, abilities and other characteristics (KSAOs). One of the notable examples is found in Rayon et al. (2014).
- xAPI – Experience API[3] is a specification for education technology that enables collection of data on the wide range of experiences a person has (both online and offline). xAPI records data in a consistent format about an individual or a group of individual learners interacting with multiple technologies. The vocabulary of the xAPI is simple by design, and the rigor of the systems that are able to securely share data streams is high. Besides regulating data exchange, there exists a body of work about using xAPI for aligning the isomorphic user data from multiple platforms. For example, Bakharia et al., (2016) discusses an example of aligning activity across multiple social networking platforms, and also provides code and data snippets.

Once one standard is chosen, it is easy for a researcher or practitioner to convert the data in another standard as needed. By committing to a data standard, the user can leverage the unique capability of the database while also prescribing structured commitments to incoming data so that robust, reliable

[2] https://www.imsglobal.org/aboutims.html
[3] https://xapi.com/overview/

processing scripts can be built. It is also possible that new standards will be developed in the future to facilitate the data exchange for new purposes.

2.3 Standard Data Exchange Model and the Statistics Domain

The National Information Exchange Model (NIEM) has been working on developing a Statistics Domain for the larger Federal community (NIEM Statistics Domain Kicks-off, 2020). NIEM has a very succinct value statement for standardizing information for board exchange that is worth noting: "NIEM is a common vocabulary that enables efficient information exchange across diverse public and private organizations. NIEM can save time and money by providing consistent, reusable data terms and definitions, and repeatable processes.[4]" It would be advantageous for NCES to take the lead in the educational component and think broadly across the community of stakeholders including other government organizations that need these data assets.

2.4 Rich Context and Metadata

When researchers collect and curate data assets, the ability to capture their tacit knowledge is critical. Beyond just a standard data dictionary, rich context about the data will assist in the creation of knowledge graphs and discover or connect research using a data asset. Understanding standard interpretations of variables, how it was processed, what it means, what is observed, or anything about the design of the experiment that would be important when making inferences from the data or drawing conclusions is vital for the researcher. Historical knowledge of any variable name changes, why it was done, and what can be done to preserve data linkages for longitudinal studies is one more example of why this is critical to build into the data asset early.

III. SUGGESTED FORMATS FOR OTHER PROCESS DATA RELEASE

The previous section provided our recommendations about selecting a data standard for release of process data to researchers who prefer access to the data in its rawest form. However, it is likely that many, if not most researchers, regardless of their ultimate goals, will have a need for a common set of summary information from the process data that must be obtained by processing the raw data files. Part of the panel's work was to identify useful summaries that NCES could provide, so that individual researchers would not need to duplicate each other's efforts. Besides improving efficiency, we believe that having all researchers work from the same set of summarized data can improve reproducibility by ensuring that concept definitions are operationalized uniformly.

In this section, we describe two forms of data release, one for researchers whose research questions requires access to original data analyses on the process data, but who do not have the resources or desire to process the data in its rawest form. The second form is for those researchers or practitioners that may have relatively straightforward and specific requests for process data summaries, but do not have the facilities or interest in handling the any microdata files themselves.

[4] https://www.niem.gov/

3.1 Release of Pre-Processed Microdata Files

Since some researchers will conduct analyses on items and others will focus on test-takers, we recommend that two pre-processed files be prepared; one focused on items and one on examinees. These two types of pre-processed summary files should be produced for each content area (e.g., reading, writing, mathematics, science, etc.) for which process data is released, but the variables contained in each will differ somewhat, depending on actions required by examinees. For both item and examinee files, summary information comes in two forms: background information and summary statistics computed from the process data itself. Tables 1 and 2 provide listings of variables that we recommend be included in the item and examinee files for all content area assessments. Section (a) of each table lists the background variables, and Section (b) lists the process data summary variables.

First we note that when the process data are released, it will contain data for released items only. Our panel believes this is a reasonable approach, since in addition concerns about security, the process data without knowledge of the item details would be of limited utility to education researchers. For the 2017 Grade 8 Mathematics assessment, the data files will consist of process data for two blocks of items (or roughly 30 items) and for all the examinees who took both (about 3K examinees) or only one (about 25K examinees) of the released blocks.

Table 1 (a) shows that most of the item background information we are recommending be included on the file, such as item parameters and subscale, could be obtained by researchers from other NCES sources. Our goal for recommending such variables be included though is to reduce the duplicative effort each individual researcher must spend to assemble useful data.

Table 1. Suggested variables for item file*

(a) Background Variables		(b) Process Summary Variables	
Variable Name	Description	Variable Name	Description
Item ID	Provides key to link the item to examinee file	Summary statistics for time	Mean, SD, Q1, Q3, max and min cumulative time spent on item by examinees
Position	Position of the item within the block	Summary statistics for visits	Mean, max, min # of times the item was visited by examinees
Item parameters	IRT parameters, p-value	% answer change	% of examinees who changed their answer one or more times
Subscale	The subscale (e.g. algebra) to which the item contributes	% omitted	% of examinees who reached but did not answer the item
#words, #images, #tables	Number of words, images, and tables in the item statement	% not reached	% of examinees who did not reach the item
Tool indicator (separate variable available for each tool, if more than one)	Presence or absence of tools available to examinee (e.g., calculator for mathematics assessment)	% tool use (separate variable for each tool, if more than one)	% of examinees who used the available tools for the item (e.g., calculator for mathematics item)
Correct response	Correct answer for MC	Responses summary	For MC items, the % of examinees who selected each option
DoK	A measure of the depth of knowledge/cognitive level at which the item was intended		

* In this file, the record is an item in the assessment, and all variables are descriptive of the item.

The details of some of the recommended variables will differ by content area. For example, for mathematics assessments, the availability of the calculator would be included as a *Tool indicator* variable, but would not be for other content areas. The indicator will record whether or not the item required the test-taker to interact with a multi-media file as part of the stimulus passage (e.g., listen to an audio file or view a video a file). For some items, such as extended constructed response items, the *Correct response* would not be included.

The recommended variables shown in Table 1(b) will require pre-processing by NCES staff to obtain. All the variables listed can be computed by summarizing the listed item characteristics over examinees. Though this increases the resources required to prepare the file, we believe it could provide both economic and scientific advantages. First, the variables listed are those the panel believe most researchers would want to examine. Thus each researcher would need to compute the statistics for themselves if they were not provided, resulting in redundant effort and even cost to NCES, if they were supported grantees. Second, because of the complexity of the process data, operationalizing the variables could differ from one researcher to another. Even a seemingly straightforward variable like the number of visits to an item is open to interpretation, since a decision must be made as to whether or not some minimum time must be spent on the item to count as a visit. NCES is in the best position to understand a meaningful way to operationalize each variable since they understand the way the system records the examinee inputs. By providing the summary statistics, along with meta-data that describes the way the variable is computed, NCES can help improve the reproducibility of the findings made from the data. Of course, if a researcher prefers to use a different definition, he or she may do so by accessing the raw process data file.

The purpose of the item summary information in Table 1(b) is to provide researchers with a normative measurement scale for examinee behaviors against which subgroups of examinees can be compared. These could be used in research similarly to how average scale scores and achievement levels allow comparisons for cognitive measures. The panel discussed whether or not the summaries should be calculated as if they are estimates of population parameters as the cognitive summaries are; i.e., whether they should take the complex sample design into account by incorporating weights. We did not reach a definitive conclusion on this issue, as we believe the research question being addressed will determine whether the appropriate comparison is a population level one or not. More discussion of this issue is included in Section 5.

Table 2(a) and (b) display the background and process summary variables we recommend for the examinee file. The records in the examinee file refer to individual examinees and include information both about their cognitive performance and their process behavior. The examinee background variables include some school characteristics (such as school type and % of students receiving free and reduced-price lunch) and some person level characteristics (such as IEP status and estimated performance level). We also recommend that information about the complex design be included as a background variable in the file. This will include the sampling weight, and might also include stratification and cluster ID variables, depending on the design.

Table 2. Suggested variables for examinee file*

(a) Background Variables		(b) Process Summary Variables	
Variable Name	Description	Variable Name	Description
Demographic variables	Race/ethnicity/gender	Item Path (k_i component array, where k_i=# of item visits by i^{th} examinee)	Listing of item #'s visited by the i*th* examinee, in order
SES measures	Indicator of eligibility for free/reduced-price lunch for student and proportion in school	Time path (k_i component array)	Time spent on each item visited by the i*th* examinee, in order
Disability information	Indicator of IEP status of examinee and nature of disability	Tool use for each type of tool; e.g., calculator, read-aloud, etc. (k_i component array)	Indicators of whether a tool was used during each visit to the item
ELL status	Indicator of ELL status of student	# Omitted	The number of items which the examinee visited but did not record a response.
Accommodation indicators	Indicator(s) of accommodation(s) received for assessment	# Not reached	The number of items which the examinee did not visit
Performance level	Examinee's estimated performance level, based on all items taken	Total time	The total amount of time the examinee spent before leaving the block
Total points	Examinee's total # of correct answers in the block	%-ile time	Examinee's percentile rank in Total time.
School code	Code indicating school so that students in the same school can be identified across the file	Answers (*n* component array, where n_i=# of item in the block)	Examinee's answer to each item
School demographic variables	Type of school, state or TUDA	Points/score (n component array)	Examinee's score for each item
Complex sample design information	Stratum, Cluster, and Sampling weight	Item time (n component array)	Total time examinee spent on each item
Block(s)	Code indicating which block or blocks of the released data the examinee received	Item visit count (n component array)	Total number of visits examinee made to each item
		Item tool use for each type of tool; e.g., calculator, read-aloud, etc. (n component array)	Indicators of whether examinee used tool on any visit to each item

*In this file, the record is an examinee who received the process data block in the assessment

For the examinee file, two categories of process variables are included in Table 2 (b). The first category summarizes the examinees' experience performing the full set of items in the block. The second category provides information specific to each item.

The first three variables in the first category of process variables are arrays, whose identical length (which we will denote by k_i) is the number of visits made to any item by the i^{th} examinee. The first variable listed in Table 2 (b) is Item Path, a k_i -component array indicating the order in which each item

was viewed by examinee i, including revisits to a given item. For example, (1,2,3,2…, 10) would indicate the examinee visited item 1 first, then item 2, then 3, then returned to item 2, and visited item 10, just before exiting the block. Because examinees will take different paths through the block, the array length k_i will vary over examinees. For some content areas, examinees may interact with objects other than items, so that the possible components of the array may include them as well. For example, a reading assessment item path may include reading passages as components. The second variable listed in Table 2 (b), called Time path, is also an array of length k_i, whose components show the amount of time the examinee displayed each item during the visit to that item. The third variable shown is Tool use, which is a generic name for several possible k_i-component arrays of indicators (e.g., 0/1) of whether or not a specific tool, such as a calculator, was opened by the examinee during the visit to the item. The content area of the assessment will determine how many Tool use arrays are required.

The next six variables listed in Table 2 (b) are single summary statistics describing features of the examinee's path through the block. These include the numbers omitted and not reached, the total amount and %-ile of time spent in the block, as well as an index of speededness; i.e., whether the examinee had sufficient time to finish the block. A consistent rule based on examinee behavior will need to be established as an indicator of this characteristic.

Finally, the last five variables in Table 2 (b) provide data for each individual item. They can also be thought of as arrays, but with identical length n for all examinees, where n is the number of items in the block. These five variables record the answer and score received for each item in the block, the total time and count of the number of separate visits the examinee made to each item in the block, and an indicator of whether or not the examinee used each available tool (such as the calculator) on any visit to the item. Clearly, these variables are simply functions of the path array variables, but we believe are preferably computed and provided to the researcher by NCES.

The panel noted that the type of summary information needed by researchers and the conceptual structure above would also be appropriate for the NAEP reading tasks, although with different definitions for components of the item path and tools available to the examinee. Writing tasks, however, will require notably different summaries for its process data than other content areas. Additional considerations on writing data are provided below in Section 4.

3.2 Release of Process Data Summaries

The pre-processed files discussed in the previous section will make it much easier for researchers to extract information from the process data files. However, even these files will require technical skills and sensitive data handling facilities that not all who might be interested in the data have readily available. We also believe that practitioners, such as school district personnel whose primary purpose is not research but rather decision making for school systems, may find that the process data can provide useful insight into student performance. These potential data users would benefit from being able to access basic summaries of process data for subgroups.

Therefore, the panel recommends that a tool similar to the NAEP data explorer be developed to provide these summaries for process data. The NAEP data explorer is designed to provide comparisons of student assessment scores for demographic or geographic subgroups. The user can choose to compare

scores for a composite scale or specific subscales. We recommend that the process data explorer should contain summaries for released items only, and should provide summary statistics at the item level. These summaries should include some of the same information discussed in the description of the item file in the previous section. Examples of useful summaries would be ones describing the time distribution and the tool usage by students on each item.

The panel believes that it is important that the user have the ability to compare these statistics across student subgroups at the item level. Without the context of the item, the summaries will not provide much insight into student behavior. If the number of released items available becomes so great that this is not feasible, then we recommend that the summaries be provided for a subset of items that contain exemplars of item types (e.g., multiple choice, constructed response), difficulty level, and scale.

IV. PROCESS DATA FOR WRITING

The charge of our panel was to consider the first release of process data to researchers, which was the 2017 Grade 8 Mathematics Assessment. However, the panel did discuss some issues related to release of process data from other content areas. Most other assessments have items similar to that of the mathematics assessment (e.g., multiple choice and short constructed response items), so the general formats discussed in Section 3 will still apply. However, there are some content areas whose items require very different test-taking behaviors by examinees. Examples of these are NAEP's Technology and Engineering Literacy (TEL) and Writing assessments. Our panel gave some consideration to the Writing assessments and our recommendations are discussed in this section.

Writing in and of itself is a process and every keystroke and selection made by the examinee in NAEP writing is recorded. Accessing the writing process data would provide an opportunity for the researcher to analyze the writing portion of NAEP beyond just the scores that examinees receive and to deeply examine the process underlying the student work as well as the complexity of the language (e.g., Deane & Zhang, 2015) they use. The panel notes that there are important considerations in the analysis of NAEP writing that will affect the kinds of analyses that can be performed on the data.

Like the mathematics data, the panel recommends providing summary files which would include some of the same information about the items and the examinees as shown in Table 1 (b) and Table 2 (b). These would include such variables as mean time on each item and indicators and summaries of tool use. We also recommend some new summary variables. For example, in the examinee file, new variables specific to writing fluidity would include mean words created per minute, and percent of writing that is deleted.

However for the actual writing keystroke data as discussed in Section 2, care must be used to prevent exposure of the raw process data. Writing data may contain Personally Identifiable Information (PII). For example, NAEP prompts may ask a student to convey a particular experience which can cause some students to use personal information from their own experience. As such, wider distribution of this data to researchers may be problematic. The committee recommends two solutions.

First, writing data could be distributed using a masked version of the writing process data which removes all specific word and letter information. It would still retain information about the time each word was created and actions such as deletions representation. This would require pre-processing the data to convert it to provide just information about the start time and end time of each word or sentence created, use of punctuation, and deletions.

A second alternative which would be preferable from the panel's point of view is to provide the full writing process data after doing a manual review to remove any PII or to release the data with strong restrictions to a limited number of researchers. Prior releases of large-scale writing data have used anonymization of key entities. The Kaggle Automated Scoring Assessment Prize performed a post review of the writing data to remove any PII using a combination of a named entity extraction program and human reviewers. This resulted in sentences being converted from one such as *"I attend Springfield School..."* to *"...I attend @ORGANIZATION1"* and *"once my family took my on a trip to Springfield."* to *"once my family took me on a trip to @LOCATION1."* This approach would provide researchers with much richer information about the creation and expression in the writing, for example showing the complexity of vocabulary, the proper use of grammatical constructs, and the quality of argumentation and supporting details. This approach would open up a range of research questions that could be investigated; for example, do students initially outline their main arguments and then fill in details or do they tend to process through the essay in a linear order? Further consideration and consultation with experts on writing research is advised in order to inform the summary information about student processes employed when generating written responses.

V. EXAMPLES OF RESEARCH PROJECTS ADDRESSABLE WITH PROCESS DATA

A variety of communities use NCES assessment score data for decision-making and research. For NAEP, the main focus of this report, the data about what American students know and can do are used by school policy-makers, teachers, curriculum developers, and even the media to make decisions about how to improve learning and to provide information to the public. The assessment process data might prove useful for the same purposes if it were to become easily available and understandable. Before that is possible, however, researchers will need to determine what information in the data is most useful for comparing educational practices and identifying meaningful student behaviors. At this point, it is not even clear which behaviors those are, since the ability to study these factors has not previously been widely available. Our motivation for providing the process data in the three formats discussed in this report was to provide access to this resource for the widest possible audience to begin to address such questions.

We believe that researchers interested in enhancing teaching and learning, improving educational assessment, and even developing new psychometric and educational statistics methodologies will all find research opportunities in these new data. The goal is that these researchers will identify questions that can be addressed more effectively by combining process data with performance data than by the latter alone. Examples of such questions are whether the student's score accurately reflect their ability or not, how to improve assessment so that they do accurately measure all students' abilities, what tools

are most effective for students with disabilities or English language learners, and how better to personalize educational practices for individual students. In this section, we examine a few of these research topics.

5.1 Improving Teaching and Learning

Analysis of process data together with outcomes at the individual and group level could provide insight into strategies for successful student performance. One such goal is to identify patterns of problem-solving that do or do not correlate with high performance. If behaviors could be identified, these strategies could be investigated with experimentation to establish effectiveness, and new interventions developed. Examples of these could be to identify how high performing students with disabilities use the tools available to them, what revising strategies in writing lead to the best outcomes, and whether highlighting in reading passages are associated with improved comprehension.

The pre-processing of the raw data file as discussed in Section 3.1 will make these kinds of analyses more efficient. We anticipate that many users will focus their analyses on sub-sets of items, sub-sets of examinees, or both. As an example, a researcher interested in the effect of the read aloud support tool could focus their analyses only on those items for which a substantial proportion of examinees actually employed the read aloud tool. Providing summary information about read aloud tool use for each item would allow a researcher to efficiently identify those items that are of interest and those that are not. As another example, a researcher interested in the relationship between a specific demographic variable and the path taken to work through the items could efficiently categorize test takers into different types of paths using the Item path array without having to analyze the full process data file to define and categorize item paths.

We note that some questions requiring comparisons of outcomes for subgroups of students are best made at the population level. For example, a question such as *"Do students who use writing strategy A score better on average than students who use writing strategy B?"* require that the sample design is considered, in order to properly account for differential selection probabilities and sample correlations. For this reason, the sample design characteristics should be included in the released files as well as used in the NAEP process data explorer, as recommended in Sections 3.1 and 3.2.

5.2 Improving Assessment

Analysis of process data can help inform the improvement of future administrations of large-scale assessments. A particularly relevant example of this is test-taking engagement. As a low-stakes assessment, the validity of NAEP outcomes depends on the assumption that examinees give their best effort. Item response time and other types of process data have been found to be useful in identifying disengagement. This information could be used to address a variety of research questions, such as the prevalence of disengagement during NAEP assessments, the administration conditions and types of items most vulnerable to disengaged responding, and the degree to which NAEP outcomes are distorted by disengagement.

An even more basic question to investigate that would provide useful feedback to test developers and content experts is *"How does the typical examinee behave?"* Establishing whether they progress through

the items and the blocks as developers assume they do, and how much unproductive time they spend experimenting with tools or other activities could inform developers about the digital interface. An understanding of examinee behaviors could also inform IES whether the convention assuming incomplete items at the end of a block should be considered "not reached" instead of "omitted" is logically supportable.

Study of the item paths, tool usage, and other examinee behaviors could help IES identify, and even eventually anticipate, how and why the features of an electronic device may artificially impact examinee performance, or help them determine if the tutorial the examinees are provided needs improvement. It could also help researchers unravel unexpected results and help determine whether certain subgroups of students may be disadvantaged by computer inexperience rather than knowledge and skills. For example, if gaps between reporting subgroups in NAEP widen after the pandemic, examination of process data may reveal if any of the difference is due to improved keyboarding skills or familiarity with computer interfaces in some groups, or something else.

5.3 Methodological Development

Access to the process data provide psychometricians and others developing statistical methodologies opportunities to explore a variety of approaches to improve scoring. For example, could some process data summaries prove to be useful background variables to improve plausible value modeling? The process data have complex structure, and some of the simplifying assumptions used for the cognitive response data, such as conditional independence, may prove too restrictive. So researchers may find it useful to develop new methodologies, or to work with aggregated data, to handle dependencies in the process data.

Since so little is known about the value of information in the process data for scoring, much of the early research will likely be exploratory. The newly observable dimensions of time and tool usage in the process data make many of the conventional methods of exploring data, such as histograms, scatterplots, and summary statistics inadequate for the task. Thus development of creative data visualizations, and investigations of which are most revealing and informative to users will be valuable contributions to the process data analysis toolbox.

5.4 Future Iterations of Process Data

The recommended format for release of process data will surely change as researchers discover what information is useful and what is not. We believe that NCES should plan for frequent changes as the research community makes discoveries that we have not anticipated. For example, we have not suggested that any of summary information from the NAEP student or teacher survey questionnaires be provided as background variables in the pre-processed data files. It may be that some questionnaire items will be frequently enough used that it would be efficient to provide them to researchers. The panel suggests that NCES should develop mechanisms for early users of the process data from each content area to provide feedback on ease of use and further requests for summaries.

APPENDICES

A. References

B. Charge to Panel

C. Agenda

D. Technical Expert Panel Biosketches

Appendix A: References

Bakharia, A., Kitto, K., Pardo, A., Gašević, D., & Dawson, S. (2016). Recipe for success: lessons learnt from using xAPI within the connected learning analytics toolkit. In Proceedings of the sixth international conference on learning analytics & knowledge (pp. 378-382). ACM.

Cooper, A. (2014). Learning analytics interoperability-the big picture in brief. *Learning Analytics Community Exchange.*

Deane, P., & Zhang, M. (2015). Exploring the feasibility of using writing process features to assess text production skills (Research Report No. RR-15-26). Princeton, NJ: Educational Testing Service. http://doi.dx.org/10.1002/ets2.12071

National Information Exchange Model Statistics Domain Kicks-off. (2020, February 27). Retrieved April 6, 2020, from https://www.niem.gov/about-niem/news/national-information-exchange-model-statistics-domain-kicks

Rayon, A., Guenaga, M., & Nunez, A. (2014). Ensuring the integrity and interoperability of educational usage and social data through Caliper framework to support competency-assessment. In 2014 IEEE Frontiers in Education Conference (FIE) Proceedings (pp. 1-9). IEEE.

von Davier, A.A., Chung Wong, P., Yudelson, M., Polyak, S. (2019). The argument for a "data cube" for large-scale psychometric data. Frontiers in Education. https://www.frontiersin.org/articles/10.3389/feduc.2019.00071/full

von Davier, A.A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement*, *54*(1), 3-11.

Appendix B: Charge to Process Data Panel

Broad access to process data will require providing researchers with opportunities at two levels: access to the raw process data files at the student level and new variables created from that data. An earlier panel was held to deal with the technical aspects of creating process data files. The current panel is charged with considering the substantive issues of what specific research or policy questions should be addressable from these data. In addition this panel is asked to provide guidance on the creation of new variables that will be useful to a broad collection of researchers in the fields of education sciences, test development and psychometrics, behavioral psychology and related fields. Guidance may take the form of criteria for definition of new variables but must include specific examples. Specifically, we are looking for variables that most researchers will end up creating themselves if NCES does not include them in the data file.

Appendix C: Agenda

NCES Panel on Release of Process Data to Researchers

March 11–12, 2020 | Washington, DC

PCP Room 4080-4090

There are certain times when NCES staff will participate. These are:

Tuesday: *9:00 am – 12:00 pm session (staff invitees attend – as needed or interested)*
4:30 pm – 5:30 pm session (staff will attend as needed)
Wednesday: *10:00 am – 10:30 pm session (staff will attend as needed)*
12:30 pm – 2:30 pm session (staff invitees attend – as needed or interested)

March 11, 2020 (PCP 4080)

8:30 a.m.	Arrival and building security
9:00 a.m.	Welcome and NDA signing
9:15 a.m. – 10:00 a.m.	NCES expectations and discussion with the panel
10:15 a.m. - 11:15 a.m.	AIR process data demonstration with the panel
11:15 a.m. – 12:00 p.m.	Discussion
12:00 pm. – 1:00 p.m.	Lunch (on your own)
1:00 p.m. – 4:30 p.m.	Panel Executive Session
4:30 pm. – 5:30 p.m.	Clarification requests from NCES/AIR (as needed)
5:30 p.m.	Adjourn

March 12, 2020 (PCP 4090)

8:30 a.m.	Arrival and building security
8:45 a.m. – 10:00 a.m.	Panel Executive Session
10:00 a.m. – 10:30 a.m.	NCES/AIR responses to panel requests (if useful)
10:00 a.m. - 12:00 a.m.	Panel Executive Session
12:00 pm. – 12:30 p.m.	Working Lunch (group will leave and bring lunch back
12:30 p.m. – 2:30 p.m.	Summary Session with NCES
2:30 pm. – 5:00 p.m.	Panel Executive Session
5:00 p.m.	Adjourn

Appendix D: Expert Technical Report Panel Biosketches

Steven L. Wise, Ph.D.

Title: Senior Research Fellow, Collaborative for Student Growth, NWEA

Dr. Steven Wise has published extensively during the past three decades in applied measurement, with particular emphases in computer-based testing and the psychology of test taking. In addition, he sits on the editorial board of several academic journals and has provided psychometric consultation to a variety of organizations, including the Maryland State Department of Education, the Virginia State Department of Education, the Nebraska State Department of Education, the National Assessment Governing Board, the American Board for Certification of Teacher Excellence, and the GED Testing Service. In recent years, Dr. Wise's research has focused primarily on practical methods for effectively dealing with the measurement problems posed by low test taker engagement on achievement tests.

Alina A. von Davier, Ph.D.

Title: Senior Vice President, ACTNext (a multidisciplinary innovation unit at ACT Inc.)

Alina von Davier, Ph.D. is a pioneer in Computational Psychometrics, an emerging interdisciplinary field concerned with the application of theoretical psychometric models and data-driven computational methods for multimodal, large-scale/high-dimensional learning and assessment data. Von Davier's unique approach drives ACTNext's development of innovative solutions to challenging problems, and challenges the ways in which assessment is traditionally thought of. Her current research interests involve developing methodologies in support of adaptive learning in virtual environments that allow for collaboration, using techniques incorporating machine learning, data mining, Bayesian inference methods, and stochastic processes.

Two publications, a co-edited volume on Computerized Multistage Testing (2014) and an edited volume on test equating, Statistical Models for Test Equating, Scaling, and Linking (2011) were selected as the winners of the Division D Significant Contribution to Educational Measurement and Research Methodology award at American Educational Research Association (AERA). Additionally, she has written and/or co-edited five other books and volumes on statistic and psychometric topics. She has received significant grants and contracts as a Principal Investigator; funding sources have included the National Science Foundation, the Spencer Foundation, the MacArthur Foundation, the US Army Medical Research, and the Army Research Institute.

Prior to leading ACTNext, von Davier was a senior research director at Educational Testing Service (ETS) where she led the Computational Psychometrics Research Center. Previously, she led the Center for Psychometrics for International Tests, where she was responsible for both the psychometrics in support of international tests, TOEFL® and TOEIC®, and the scores reported to millions of test takers annually.

Von Davier is currently an adjunct professor at the University of Iowa and Fordham University, and the president of the International Association of Computerized Adaptive Testing (IACAT). She currently serves on the board of directors for the Association of Test Publishers (ATP), and she is also a member of the board of directors for Smart Sparrow and of the advisory board for Duolingo.

She earned her doctorate in mathematics from Otto von Guericke University of Magdeburg, Germany, and her master of science degree in mathematics from the University of Bucharest, Romania.

Michael Russell, Ph.D.

Title: Professor, Measurement, Evaluation, Statistics, and Assessment, Boston College

Dr. Michael Russell received his Ph.D from Boston College. His scholarship focuses on validity theory; race and quantitative methodology; innovative uses of computer-based technologies and applications of Universal Design to enhance educational testing and assessment; large-scale assessment and test design; computer-based testing; and Accessible Portable Item Protocol (APIP) Standards and assessment interoperability standards. He was the founder and Chief Editor of the *Journal of Technology, Learning and Assessment* and provides technical support to several state assessment and accountability programs.

Peter Foltz, Ph.D.

Title: Vice President, Pearson's AI & Products Solutions / Research Professor, University of Colorado's Institute of Cognitive Science

Dr. Peter Foltz's work covers artificial intelligence and uses of machine learning and natural language processing for educational and clinical assessments, large-scale data analytics, reading comprehension and writing skills, and 21st Century skills learning. He has served as content lead for framework development for several Organisation of Economic Cooperation and Development's (OECD) Programme for International Student Assessment (PISA) assessments, including the 2018 Reading Literacy assessment, the 2015 assessment of Collaborative Problem Solving, and a new assessment of reading literacy for developing countries. He has served as guest editor for a number of journals including *International Journal of AI in Education* and *Discourse Processes* as well as co-editor of the recent *Handbook of Automated Assessment.* He previously worked at New Mexico State University, Bell Communications Research, University of Pittsburgh's Learning Research and Development Center, Yale University, and the Harvard Institute for International Development.

Brock E. Webb, M.S.

Title: Senior Information Technology Policy Advisor Office of Management and Budget (OMB), Statistical Science and Policy (SSP), Office of Information and Regulatory Affairs (OIRA)

Brock Webb, is currently on special Detail to the Office of Management and Budget as a Senior IT Policy Advisor. In his permanent post at the US Census Bureau, he led the Cloud Program which has transformed the way Census acquires and consumes IT. He also drove the Census acquisition strategy and FedRAMP sponsorship of the New York University's Administrative Data Research Facility (ADRF), which enables joint data sharing and statistical work across multiple Federal, State, and Local partners in support of the U.S. Commission on Evidence Based Policy. Prior to joining Census, Brock was the Chief Engineer for Cloud Computing in the Chief Technology Office at the Department of Defense (DoD) Defense Information Systems Agency (DISA).

Technical Expert Panel convened by National Institute of Statistical Sciences

S. Lynne Stokes, Ph.D.

Title: Senior Research Fellow, National Institute of Statistical Sciences-DC; Professor & Chair, Department of Statistical Science, Southern Methodist University

Lynne Stokes is an expert in surveys, polls and sampling, as well as in non-sampling survey errors, such as errors by interviewers and respondents. She is a Fellow of the American Statistical Association. She recently has conducted research on evaluating the accuracy of contest judges and on improving estimates of marine fishery yields by the National Oceanic and Atmospheric Administration.

She also contributes to the National Assessment of Educational Progress, or "Nation's Report Card," examining the way schools and students are selected for the large study. Stokes became a faculty at Vanderbilt University, but in 1979 began working for the U.S. Government as a statistician, first at the Patuxent Research Refuge of the U.S. Fish and Wildlife Service and then at the Census Bureau. She returned to academia in 1984, at the University of Texas at Austin, and moved to Southern Methodist in 2001. Research Interests: Surveys, Polling and Sampling, Voter Exit Polling, Sampling Methods, Non-Sampling Errors, Non-Disclosure Methodology, Measurement Error, Order Statistics, and Mark and Recapture Methods.

Nell Sedransk, Ph.D.

Title: Director, National Institute of Statistical Sciences-DC

Dr. Nell Sedransk is the Director of the National Institute of Statistical Sciences. She is an Elected Member of the International Statistical Institute, also Elected Fellow of the American Statistical Association. She is coauthor of three technical books; and her research in both statistical theory and application appears in more than 60 scientific papers in refereed journals. The areas of her technical expertise include: design of complex experiments, Bayesian inference, spatial statistics and topological foundations for statistical theory. She has applied her expertise in statistical design and analysis of complex experiments and observational studies to a wide range of applications from physiology and medicine to engineering and sensors to social science applications in multi-observer scoring to ethical designs for clinical trials.

Alexandra Brown, M.S.

Title: Research Assistant, National Institute of Statistical Sciences-DC

Alexandra Brown is a Research Assistant at the National Institute of Statistical Sciences working under the direction of Dr. Nell Sedransk on projects in education research. She holds a MS degree in Economics and is currently a PhD candidate in Survey Methodology at the University of Maryland.