

Institute of Education Sciences  
National Center for Education Statistics

## MAKING NCES PROCESS DATA AVAILABLE

NISS Technical Expert Panel • October 2019

### EXECUTIVE SUMMARY

Most National Center for Education Statistics (NCES) assessments and surveys are now conducted using electronic modes. In consequence, the data captured include not only responses but also time-stamps and click-by-click chronicles of the response process. These data offer unique insight into the cognitive processes involved in test-taking. They also have potential use for automated scoring and may be useful from a psychometric point of view in evaluating item properties. NCES is now committed to making these process data available to researchers.

Therefore NCES charged the National Institute of Statistical Sciences (NISS) with convening a panel of technical experts to advise on how to present these data efficiently, effectively and comprehensively without incurring unacceptable security, privacy or disclosure risks.

Process data comprise the time and action information as a student progresses through the assessment so that the pattern of a student's attention is evident as is the time spent with each action. Questions that are not answerable from response data alone can be studied from process data by the broad community of test developers, psychometricians, behavioral psychologists and many education researchers.

From these time/action data researchers can study student's progress through the assessment including use of tools, time spent on prompts and questions and patterns and sequences for question response and review/correction. Analysis of how time is spent during progression through the assessment may illuminate more elusive concepts such as engagement or lead to definition of new constructs. Differences in test-taking patterns and implications for performance scores can be studied for subgroups of students such as English language learners (ELL) or for students with particular needs such as dyslexia. Process data can also be used during the process of test development as psychometric properties are evaluated; and test design could specifically address new constructs or new metrics.

NAEP process data for released items will provide an initial example. Unlike complete assessment data releases, process data will be made available for individual released items combining student records over several years the item appeared in NAEP assessments.

The immediate challenge is that raw process data are extensive and not structured for analysis or direct interpretation. To date, analyses of process data (from other sources than NCES) has required creation of custom data files for each project or proposed analysis. The charge to this panel is to consider the technical issues in structuring NCES process data so that research can be conducted. *Substantive issues of what specific research or policy questions should be addressable from these data are reserved for a separate panel to consider.*

The panel's deliberations covered two broad areas: the data users' information needs and technology capabilities, and the collective concerns surrounding security, privacy, and disclosure risks. In addition the panel pointed out the need to allow for evolving future demands as research leads to new constructs and new metrics, adopts new technical approaches and creates new computational methods.

Data users (for now) can be assumed to have two levels of technological expertise for dealing with these data: those with resources to extract information from "raw data files" into custom files and those without. For the latter group, much research can be done using well-chosen *summary variables*, if NCES provides these in addition to the raw process data files. Such summary variables would be created from the raw process data to constitute a basic set of summary information (e.g., frequencies, sequences and intervals such as time allocation by item/question, frequency, type and sequence of change(s) to original response, use of available tools such as spell/grammar check).

Specific to IEP and ELL research is the need of many if not most researchers for matched subsets of non-IEP or of non-ELL students as the basis for comparative studies. NCES is best able to create these subsets efficiently and cost effectively.

Risks to be minimized include: privacy of students, also teachers and schools; security against frivolous or irresponsible data use, disclosure of item information especially for unreleased items or any other breach that would lead to loss of trust in NCES integrity in administration of assessments or loss of trust in NCES data quality.

Two risks deserve formal risk assessment to determine their levels and potential impacts. It is NCES practice not to disclose writing samples. Release of raw text process data would violate this practice because these would allow reconstruction of original writings. To be acceptable an alternative would have to allow access to process data without access to actual content. Process information on unreleased items needs evaluation of a different risk. The panel offered the caveat that this impact might take the form of revealing sufficient information to infer item content or to give information about correct/incorrect responses.

The four-tier hierarchy already in place for access to NCES data should function equally well for process data, incorporating one specific provision for raw text process data.

Panel recommendations center on assembling into a three-part record for each student:

- A. Demographic information (the same information that is available for student-level information for NAEP).
- B. Basic set of summary variables (specific for each subject-matter area to item type, when possible, otherwise specified for the particular released item).
- C. Raw process data (for released items) with raw text and writing samples redacted.

Public information (Tier 1) would include summaries from A and B that parallel the summaries of other assessment data. Protected information (Tier 2) would draw on A and B using Data Explorer (or other NCES software) in the same way that other assessment data is accessed via Data Explorer. Restricted information (Tier 3 - NCES license required) would give the researcher the option of the data files including A and B or (for researchers with the resources to extract individual custom files) the complete data files A and B plus - for released items only - C. Highly restricted information (Tier 4 - FSDRC only) would restore redacted information to C. At NCES option, based on assessment of risk, more extensive information about unreleased items might be made available at this highest level of restriction.

The panel recommends that process data for released items be released in five sets of files. The first file would be the complete population of all students. In addition two pairs of subfiles would meet the needs of IEP (Individual Education Plan) and ELL (English Language Learner) researchers. Much research in these areas requires comparison with non-IEP/non-ELL students. Therefore data for the special group of students, IEP or ELL, would comprise the first subfile. The second subfile in each pair would contain records for (preferably 3) matched non-IEP/non-ELL students.

For unreleased items, the current practice of releasing item properties could reasonably be extended to summary reporting of summary variables at least for those data that: 1) can be meaningfully interpreted in the absence of the item itself and 2) do not violate disclosure or otherwise carry an unacceptable risk.

The panel further recommended that resources on process data content and available analytic tools be made available to researchers. A primer in the NAEP process data with small examples would serve researchers accessing public and protected data (Tier 1 and Tier 2 access). Also a smaller raw data file (“sandbox”) would allow researchers (Tier 3 access) to write, test and validate extraction and analysis code.

A public repository is needed for new summary variables, constructs and metrics to enable data use to evolve as research on process data advances. At least equally important is a repository for code to allow researchers to share the (costly) efforts of writing extraction and data integration code for these process data.

[Read the Full Report](#)