# Process data, the new frontier for assessment development: rich new soil or a quixotic quest?

Provasnik, Stephen

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

**Empfohlene Zitierung / Suggested Citation:**
Provasnik, S. (2021). Process data, the new frontier for assessment development: rich new soil or a quixotic quest? *Large-scale Assessments in Education*, *9*, 1-17. https://doi.org/10.1186/s40536-020-00092-z

Provasnik *Large-scale Assess Educ* (2021) 9:1
https://doi.org/10.1186/s40536-020-00092-z

Large-scale Assessments in Education

RESEARCH Open Access

# Process data, the new frontier for assessment development: rich new soil or a quixotic quest?

Stephen Provasnik*

*Correspondence:
Stephen.Provasnik@ed.gov
National Center for Education Statistics, Institute of Education Sciences, 550 12th Street, S.W., Washington, DC 20202, USA

**Abstract**

This paper presents the concepts and observations in the author's keynote address at the May 2019 "Opportunity versus Challenge: Exploring Usage of Log-File and Process Data in International Large-Scale Assessments" conference in Dublin, Ireland. This paper recaps briefly some key points that emerged at the December 2018 ETS symposium on process data in Washington, DC, and suggests some common terminology and concepts for facilitating future meaningful discussion. It then borrows some concepts from evolutionary biologists to take a critical look at process data and some of its ethical implications before turning to discuss major applications for and challenges of using process data and logfiles. The second half of this paper then maps out the terrain in which process data emerges as a "byproduct" of the shift from paper-and-pencil testing to digital-based assessments (DBA) and looks closely at how process data interacts with the development and operationalization of assessment systems. This interaction suggests new, fresh ways of improving national and international assessment system processes, and ultimately the measurement of cognitive skills, while venturing into the new frontiers opened with process data.

**Keywords:** Process data, Logfiles, Spandrels of San Marco, By-product, Evolution, Paper-based assessments (PBA), Digital-based assessments (DBA), Data quality, Cognitive processes, High and low performers, Paradata, Inputs, Outputs, Coding, Forensic purposes, Research purposes, Performance measure

Keynote address delivered at the "Opportunity versus Challenge: Exploring Usage of Log-File and Process Data in International Large-Scale Assessments" conference, a joint collaboration between Educational Testing Service (ETS, Princeton, NJ, USA) and the Educational Research Centre (ERC, Ireland) held May 16–17, 2019, in Dublin, Ireland.

With the conversion of large-scale assessments from paper-and-pencil to computerized or digital formats, logfiles and process data came into existence, presenting the potential of revolutionary new information and insight on individual thinking and learned cognitive and problem-solving processes. Assessment developers, subject matter specialists, cognitive scientists and sundry other researchers have proposed numerous ways to capitalize on these potential possibilities and many have conducted preliminary studies or exploratory analyses with logfiles, process data, or both in order to demonstrate the value of these data as well as a concrete ways to realize some of their

great promise. For the better part of the last two decades, these proposals, studies, and analyses have been broad-ranging but have been undertaken largely independently by individual scholars, research teams, or organizations. To move beyond this initial phase of independent and somewhat haphazard efforts and begin an era of more coordinated and methodical approaches to capitalize on logfiles and process data, ETS convened an international symposium in Washington, DC, in December 2018. In collaboration with Ireland's ERC, this symposium brought together over two dozen of the leading data process researchers and large-scale assessment developers from around the word for 2 days of vigorous debate and dialogue about how to organize existing professional expertise and plan concerted efforts to realize some of the potentials of logfiles and process data.

The "Opportunity versus Challenge: Exploring Usage of Log-File and Process Data in International Large-Scale Assessments" conference, at which this Keynote address was presented, represented a follow-up collaboration between ETS and ERC to build on the symposium's conversation and conclusions. This address recaps briefly some key points that emerged at the ETS symposium and suggests some common terminology and concepts for facilitating meaningful future discussion. It explores how process data and logfile data can be used and challenges to using them. Then it attempts to map out the terrain in which process data interact with the development and operationalization of assessments in order to think critically about how to improve assessment system processes while venturing into these new frontiers opened with process data.

## Defining process data

At the ETS symposium, one of the foundational conversations among the participants pertained to the difference between "logfiles" and "process data." The two terms overlap a great deal, but they are not synonymous concepts. At the risk of oversimplifying for the sake of clarity, one can say that logfiles are everything captured in computer-based assessment—or what is now more aptly called digital-based assessment (DBA)—from the order and speed of inputs (e.g., clicks and keystrokes) to the VPN of the device used to take the assessment. Process data, on the other hand, are the empirical data that reflect the process of working on a test question—reflecting cognitive and noncognitive, particularly psychological, constructs. Process data need not be limited to what comes from logfiles; for example, they would include eye tracking information or brain imaging, such as magnetic resonance imaging (MRI) and computed tomography (CT) scans, which are not (yet) usually captured as part of DBA, or they could include outside information (e.g., observations from interviewers or testing staff as to the degree of engagement with a specific item).

Working from these definitions, it should be evident that logfiles are data *sources* for process data. One participant at the ETS symposium suggested that logfiles can be compared with video studies, in which the video images are a source of information without filters, and it is necessary to extract the relevant parts—once one has decided which parts those are. Continuing with this comparison or analogy, then, process data are the data that result from someone lassoing or coding the "relevant parts" from logfiles. A good example of this analogy comes from the Trends in International Mathematics and Science Study (TIMSS) Video Studies of the late 1990s (Hiebert et al. 2003), in which

the recordings of selected classroom lessons were coded to indicate (among many other things) the following information:

- the assigned type of class work (using as categories "whole-class work," "individual work," "pair/partner work," and "small-group work");
- the "number of words" the classroom teacher used in a ratio to the number of words students used, when talking to the whole class; and
- the proportion of the lesson spent on review of previous content versus spent explaining new content.

These data were developed based in part on a posteriori analyses of the videos and in part on existing pedagogical theories about effective class lessons. Understanding the proportion of teacher-led discussion versus student-led discussion during a lesson is clearly a relevant concept for understanding differences in classroom teaching and learning styles, but extracting from the video recordings and coding *the number of words* (vs. amount of time) used by the teacher versus students was not foreordained or mandatory but reflected a subjective choice on the part of the researchers. Likewise, parsing the lesson time into "review" and "new content" makes sense, but the lesson could just as easily have been coded dichotomously: did the lesson include "review" (Y/N)? did the lesson cover "new content" (Y/N)?

The key point here is that process data have no "natural," normative, or prescribed format. They do not exist "out there" but are constructs that need to be defined, either a priori or a posteriori, by a theory, a research question, or some other logical method. As such, it is critical to take care not to reify process data in discussions and writings but rather to continuously keep in mind (and, when helpful, make explicit) the underlying construct and intended purpose of the specific process data being considered. Likewise, it is also essential to keep looking for ways to improve our constructs or choices when creating process data.

At the end of the ETS symposium, there was general consensus that researchers need to

1. develop a systematic approach to logfiles—to answer the question of what exactly logfiles should capture, and
2. develop a theory for process data—to answer the question of how to use process data.

Attracting less attention at the symposium was the need to develop guidelines and standards for how to convert logfiles into process data. This is just as vital a point because assessments, specifically DBA, are in the midst of evolving or perhaps co-evolving with process data. Given this reality, it can be helpful to borrow some concepts from the study of evolution to contextualize the current development of DBA and process data.

## Looking critically at process data with concepts from evolutionary biologists

The first of these concepts is that of the "spandrel," which literally is the type of space formed by two adjacent arches supporting a circular dome. The term was borrowed from architecture by Stephen Jay Gould and Richard Lewontin in their now famous article "The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme" (Gould and Lewontin 1979), to provide a name for something that has emerged through evolution but which was not selected for—a feature that was a "forced move" or byproduct of characteristics that were selected for because the selected characteristics conferred a relative advantage for survival. Gould and Lewontin used the famous spandrels of San Marco to illustrate this concept because the triangular spaces (and the decorative features that adorn them) are made possible because of the existence of the spandrels formed by arches and a dome above, but they are not what were intended—the intention was to have arches and a dome; the spandrels were resulting by-products (Fig. 1).

One can posit that logfiles are, in this sense, "spandrels" resulting from the choice to move from paper-based assessments to computer-based assessments and now digital-based assessments. Process data are possible because of the existence of this spandrel—but process data themselves are an intentional use of the logfiles (as such they are not the spandrel but the opportunity made possible from it). The concept of the spandrel helps underscore the fact that DBA and logfiles were not purposefully developed *so that* process data could be extracted nor to gain insight into how students think and process information; rather logfiles were a fortuitous outcome of the conversion to DBA. As a result, there is no intended usage for logfiles or for process data, and to move forward in any concerted effort it is necessary to explore new territory and figure out where we stand, what we can do, and where to go.

The second evolutionary example that is illuminating to consider is the development of the ability to speak in humans. The range of sounds humans can make, while continuing to breathe and keep the air pressure in their lungs constant, are a direct result of the

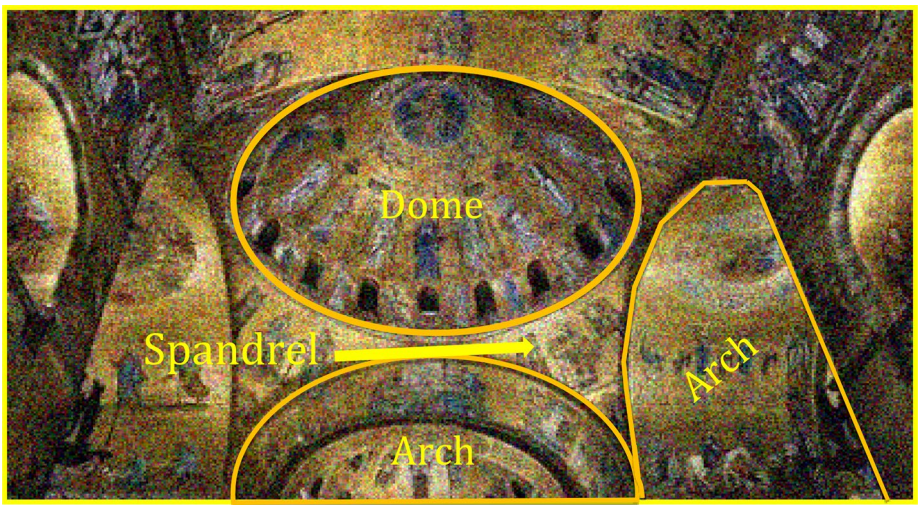


**Fig. 1** Spandrels of San Marco. Image showing a "spandrel," the space formed by two adjacent arches supporting a circular dome.

fact that humans have evolved an extraordinarily complex and risky system to speak. When humans evolved into *homo sapiens* over hundreds of thousands of years, the "voice box" or larynx (which is name for the same physical feature in primates) dropped further down the trachea to make room for the tongue as evolving human mouths and jaws protruded less and less (and human's ability to vocalize precisely increased more and more). The trade-off for the greater control and precision over the sounds that humans can make as a species, is that it is physically possible for food to go down the wrong "pipe," and individual humans can, and some do, accidently choke to death. Dogs (and other mammals), in contrast, can bolt food and have no risk of choking because mammals evolved in a way that food can only go down the esophagus. In giving up this evolutionary safeguard to make possible a broad range of sounds and speech, it became physically possible for humans to choke.

This second evolutionary development is instructive because it seems inevitable that logfiles and process data will evolve to provide more and more precise information, which is wanted, but, as they evolve, it should not be forgotten that a trade-off for that information may be that it creates the possibility for some individuals to end up accidentally paying a price for this improvement—to metaphorically choke.

The point of these two evolutionary comparisons is to underscore both that the use of process data is in its earliest stages and that its use, along with DBA generally, will evolve in many ways. The table below outlines the contours of this evolution (see Table 1). What is important to remember is that some of these ways will be intended and some not; some will confer advantages for the field but some may create new hazards for individuals.

## Major applications and challenges of using logfiles and process data

Returning to the discussions at the ETS symposium, some of the most helpful dialogue was around practical uses for logfiles and process data. The uses that seem the most beneficial (and benign) at the moment are diagnostic or forensic applications in item development and test design improvement. A wide variety of these forensic applications exist. These include using logfiles and process data to improve data quality, for example, by

**Table 1 Ongoing evolution in assessment**

| | Past | Present | Future |
|---|---|---|---|
| ***Item development*** | Labor intensive | Labor intensive | Automatized |
| ***Item types*** | Generic | Enhanced | Real-life |
| ***Test design*** | Static | Semi-static | Data-driven |
| ***Test assembly*** | Labor intensive | Semi-automatized | Automatized |
| ***Accessibility*** | Limited | Universal design | Adaptive |
| ***Timing*** | Not measurable | Measured | Used |
| ***Pathways*** | Not observable | Observable | Modeled |
| ***Validity*** | Content/core-based | Construct based | Process based |
| ***Feedback*** | Summative | Summative | Diagnostic |

Prepared by and used with permission from Ruhan Circi

1. enhancing understanding of how items function and what characteristics or variables make items more difficult or more reliable;
2. distinguishing among "missing" answers which are truly "not reached" or "not administered" (never seen), which should be "omitted" (seen, taken time over, but ultimately skipped), and which are "not attempted" (seen, but no time taken before being skipped); and
3. identifying student guessing or cases that are outliers, which may indicate possible cases of cheating, or cases of programming error.

In the National Assessment of Educational Progress (NAEP), logfiles are already regularly used to examine how items function. For example, NAEP has looked at whether students read the test questions first or read the reading passage first—logfiles showed that most students spend time on the passage first before going to questions. Similarly, logfiles have been used to examine whether students do what is expected by item writers, for example whether they go back to the passage to answer questions, especially when items direct students to look at a particular section of text (e.g., "how is the word 'royal' used in the passage?", which, when one looks back at the text, says "something was 'a royal mess'").

This sort of interaction between students, questions, and parts of an item can be neatly seen in a visualization that NAEP has posted online which presents process data from a NAEP 2017 grade 4 item "Five Boiled Eggs."

Open "See time lapse visualization" on https://www.nationsreportcard.gov/reading_2017/sample-questions/?grade=4 under Grade 4 Sample Reading Questions.

This visualization depicts sampled students (represented by blue dots), test questions, the different pages of text that students view, and each student's individualized 30-minute experience (through these pages and questions) in little more than 2 min. It visually summarizes student patterns of reading, answering items, and reviewing their answers. This is a multi-dimensional display of the collected process data that clearly and simply demonstrates that some students read the text faster than others, some slower, but most read at about the same pace, follow the same orderly steps of reading all the text and then each question, in their given order, giving rise to an average pattern that is what we would expect, with a few outliers. Despite its clarity, though, what exactly is one meant to do with this display or conclusions drawn from it? It is nothing more than a visual display over time of a more or less normal statistical distribution.

Yet another real-world example of the diagnostic and forensic applications of logfiles and process data comes from the OECD's Programme for the International Assessment of Adult Competencies (PIAAC). For PIAAC, which includes a computerized assessment, coders have programmed the automated machine scoring so that if a participant spends less than 5 s on an item, it is counted as "not administered."

NAEP has also used logfiles to examine items skipped by large numbers of students and found at least once that its system was not recording student answers properly. That is to say, what appeared to be an item that large numbers of students "skipped" was actually the result of a glitch. By going into the logfiles, NAEP was able to recover

student responses that had actually been entered but not "saved" or properly recorded by the DBA system.

NAEP has even begun to check logfiles and extract process data in real time so that adjustments/changes can be made to the assessment system while it is in the field so as not to lose an entire item due to error.

All of this is currently being done, but, as one might rightly note, these examples have made use of logfiles more than process data. However, the use of log files and process data for diagnostic and forensic purposes is in its infancy: the use of both will continue to grow and, before much longer, both will need standards and guidelines to ensure consistency across assessments in terms of item development, the coding of "missing" responses, and flags for fraud.

Besides such diagnostic uses, logfiles and process data hold out the prospect of advancing many research agendas. The great majority of these fall into the category of research into understanding respondent behaviors and cognitive strategies. Such research seeks to

- improve teaching and learning with specific information on how different students think/perform,
- better understand factors that distinguish between high- and low-performers, or expert from novice strategies, or
- better understand the relationship of motivation and performance.

Again, some examples of work done in this area are available from studies with NAEP data (Bergner and von Davier 2018; National Center for Education Statistics 2020). The figures [below] plot process data from a NAEP 2016 $8^{th}$-grade writing assessment pilot study (see Fig. 2). The horizontal axis in each of these figures shows time elapsed, while the vertical axis shows the number of characters typed by a student. The slope shows the speed at which a student is composing a written response to the test prompt.

The first student (shown in red) composed her response at a pretty steady rate with only one moment of deleting and rewriting. The second student (shown in green) had a slower rate and three episodes of significant deleting and rewriting. The third student (shown in yellow) spent almost half his time without typing anything and then wrote at a steady rate to the end of the time period. The total amount of text composed was markedly less than the first two students, indicated by the lower level reached on the vertical axis. This student had no episodes of significant deleting and rewriting. The last two show a slower rate of composition and again an overall lower level reached on the vertical axis indicating a much shorter response than the first two students. The fourth student (shown in blue) had no major episodes of deleting and rewriting but had some minor deletions and rewriting. If one puts all of these students together on a single graph, it is evident that they represent markedly different patterns and amounts of writing.

Considering these four different patterns, which students got full credit or the most points for their work? The answer is that they all received full credit (6 out of 6 possible score points). All therefore represent high-performing students. This example

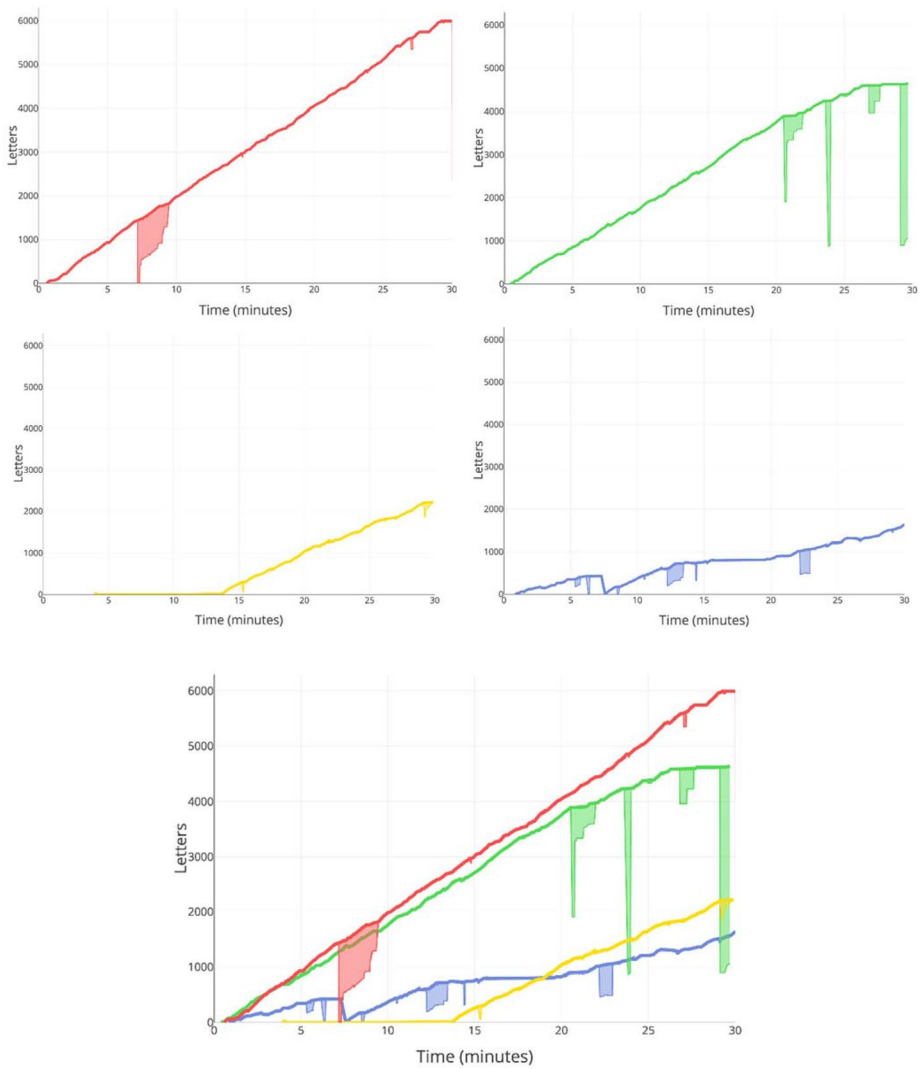


**Fig. 2** Essay Length Plotted by Writing Time, Shown for Four Students Individually and Combined. The graphs plot process data from a NAEP 8th-grade writing assessment pilot study. The horizontal axis in each graph shows time elapsed, while the vertical axis shows the number of characters typed by a student. The slope shows the speed at which a student composes the written response to the test prompt. The first student (shown in red) composes the response at a pretty steady rate with only one moment of deleting and rewriting. The second student (shown in green) has a slower rate and three episodes of significant deleting and rewriting. The third student (shown in yellow) spends almost half the response time without typing anything and then writes at a steady rate to the end of the time period. The total amount composed is markedly less than the first two students, indicated by the lower level reached on the vertical axis. This student has no episodes of significant deleting and rewriting. The last two show a slower rate of composition and again an overall lower level reached on the vertical axis indicating a much shorter response than the first two students. The fourth student (shown in blue) has no major episodes of deleting and rewriting but has some minor deletions and rewriting. If all student responses are combined on a single graph, we can see that they represent markedly different patterns and amounts of writing. From the National Assessment of Educational Progress (NAEP) 2016 8th-grade writing assessment pilot study

provides a good illustration of there being no singular "correct" strategy in writing (or indeed probably any of the other domains) for high performers.

This is an important point to make, yet for research into understanding respondent behaviors and cognitive strategies, it suggests that finding patterns of successful performance will not be a simple task.

Additional challenges emerge in another example of analysis of NAEP process data from a NAEP 8th-grade mathematics assessment, as shown in this video (Additional file 1).

[View https://youtu.be/IGj51Oj3eu4.]

This video suggests that patterns can be sorted out between high performers and low performers and that it is possible to quantify these patterns. However, there are two problems this video does not address:

(1) The patterns in themselves are not particularly meaningful, especially given that students' cognitive processes are not clearly represented (nor were they defined beforehand to be sure that they were captured), and
(2) students who took this assessment had calculators and scrap paper on hand for their use even though they input their responses into the DBA assessment with the drag-and-drop interface. Knowing this context undermines any conclusions that one can draw from the drag-and-drop order in this item's carefully cataloged and tallied sequences.

Taking time to carefully review these NAEP findings is worthwhile because they represent considerable amounts of excellent work; and yet they can have an Ozymandias quality ("Look on my works, ye Mighty, and despair!") because they can leave one with a dispirited sense as to where we stand and tempt one to think that the great and glittering promises of process data may be a quixotic quest into quicksand.

Here are just some of the challenges:

- What is the proper amount of time to program a system to treat an item as "not administered"?
- What is the right amount of time to flag a response as guessing?
- PIAAC's "less than 5 s" seems reasonable but someone could reasonably say why not a threshold of 4 or 6 s?
- Should the amount of time be variable and depend on the individual's reading speed or in some other way be individually tailored?

Peeling back layers around student "knowing," researchers may someday reach the point where it seems more appropriate to code "correct" student responses differently depending on whether a respondent demonstrates familiarity with the subject or rote knowledge as opposed to answers the item without prior experience and works out the correct answer. If one adopts such a practice, then would it be necessary to code "wrong" student responses in different ways depending on whether (a) a student attempted but failed to get the right answer versus (b) made a totally random guess? How far down this road can one go before one unravels some of the assumptions of item response theory (IRT) scaling? or will IRT evolve also, such that

someday we will have multiple types of "corrects" and multiple types of "incorrects" scaled together?

Before getting too bogged down in this type of quicksand, a move to some firmer ground seems in order. One way to do this would be to take a systematic approach to understanding what is involved in working with logfiles and process data and what is necessary for making good use of them.

## A systematic approach to seeing where process data can improve assessments and measurement

Such an approach can begin by mapping out, in the abstract, an assessment system's processes, from test development through system operationalization, both before and after DBA. Looking at an assessment system's processes before DBA (shown in Fig. 3), one finds that the formal start of an assessment system is with the creation of a framework that defines the domain of interest that is to be assessed. Once a framework has been developed, items are "written" to that framework, items are piloted, and then field test booklets are created. These booklets then are administered in a field test—the start of the operationalization of the assessment system. The field test data are then scored, item stats and item parameters generated and reviewed, and the final item pool selected—out of which final booklets are made for the main study. After the main study data collection, items are scored and IRT scaling and weighting are applied. Lastly, a final dataset is produced and analyzed to produce a report, and then it is released as the official dataset.

With the shift to digital-based assessments, the same basic assessment system persists but with several new added steps or processes that are the byproducts of the digital shift—the field's "spandrels" (see Fig. 4).

These additional processes (shown in gold) begin with the need for coders to take the items that have been written to the framework and render them digitally (typically

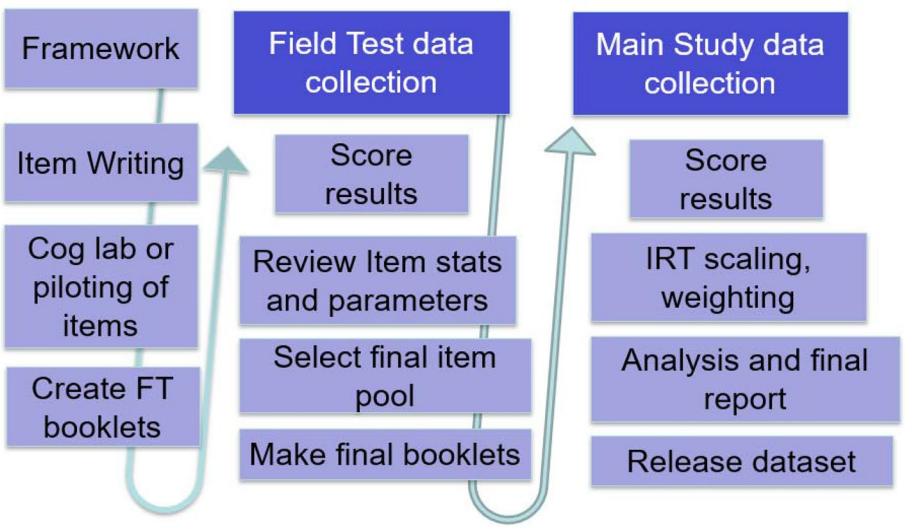


**Fig. 3** Assessment system processes before DBA. Graphical representation of an assessment system's processes, from test development through system operationalization, before DBA. Author's own work

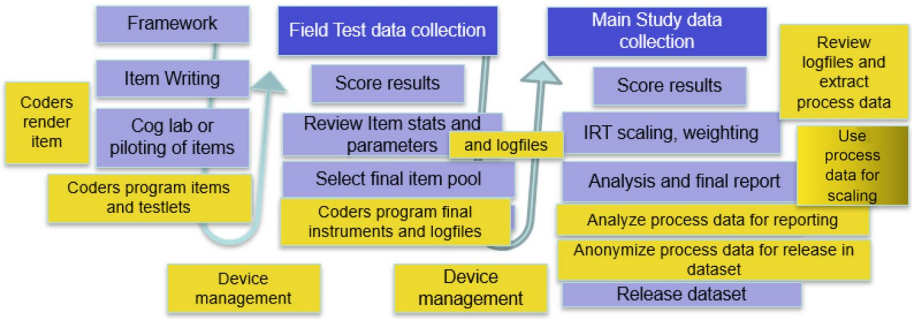


**Fig. 4** Assessment system processes with DBA. Graphical representation of an assessment system's processes, from test development through system operationalization, with new steps that are the byproducts of the shift to DBA. Author's own work

in HTML or some WYSIWYG format). Next coders program the items and testlets into the data collection system—this is governed by the IMS Question and Test Interoperability specification (QTI) and produces log files. Device management is necessary now in this digital world. After the field test data collection, log files can be reviewed along with item stats and item parameters to help select the final item pool. Coders again need to program the final instruments. After the main study data collection, it is necessary to review the log files and extract process data. These process data can be analyzed for reporting purposes and they themselves need to be anonymized if they are going to be released in the data set. One other potential use is that the process data can be used for scaling. (Note that in the figure this step is in a slightly different color font because at this point this is not typically done, though the example of PIAAC's programming of items that were viewed for less than 5 s as "not administered" is a first step in this direction.)

## Exploring the terrain's features

If this approach's mapping of the new processes in an assessment system based on DBA can be accepted, then it is makes sense to begin to look at the inputs and outputs associated with the new overall DBA system. To do this, however, it is necessary first to identify these process data inputs and outputs (see Fig. 5).

The first input in this DBA system is the need for a theoretical understanding of cognitive processes that are involved in the domain defined by the framework. Although not always traditionally done, frameworks should expound on the cognitive (and, if any, noncognitive) processes involved in the knowledge base they define.

Next, it is necessary to define the cognitive (and any noncognitive) processes that are intended to be measured by the assessment's items. This is to say that item writing itself should expand beyond defining correct and incorrect answers and justifications for distractors to include identifying and defining all cognitive processes that are to be explicitly measured.

Then, it is necessary to have deliberate coding processes that are targeted to code what is intended to be measured during the assessment.

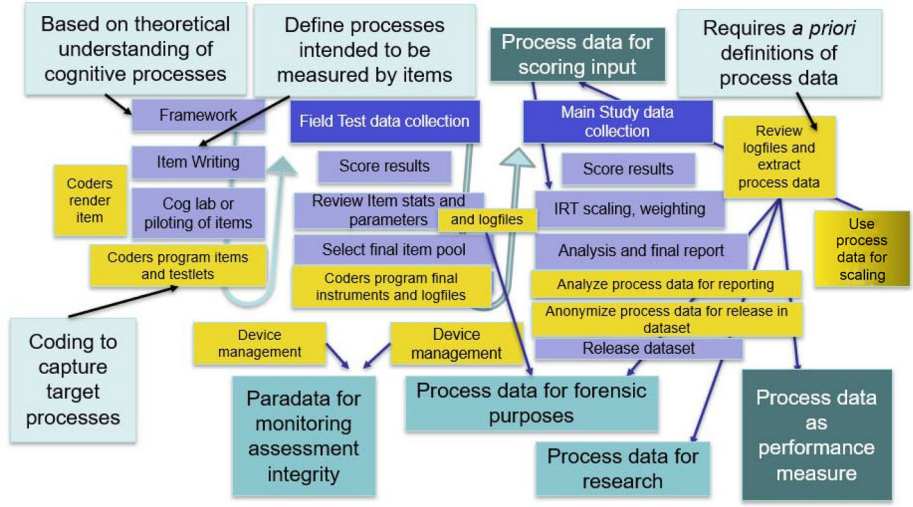


**Fig. 5** Process data inputs and outputs. Graphical representation showing test development inputs and system outputs related to process data in a DBA system. Author's own work

Finally, it is necessary to have definitions, preferably a priori definitions, of process data so that it is clear what should be extracted from the logfiles and so coding can be efficiently done to support this.

Turning to the outputs or uses of process data, the first and most overlooked type of process data—perhaps because it is typically called "paradata"—are those data derived from the use of the devices themselves that administer the assessment. These devices can provide information to identify problems with the delivery of the digital assessment as well as monitor for possible fraudulent uses of the device or cheating on the assessment.

Next, as mentioned earlier, process data can and is already being used for diagnostic or forensic purposes—both after the field test data collection and after the main study data collection.

Process data for research purposes is the area that gets most attention, as it holds out some of the greatest promise for process data. Thus, if the box for this output were to convey its relative weight, it should be larger.

Finally, there are two aspirational uses of process data that represent venturing into the new "process" land:

- process data to be input in the scoring process, and
- process data itself being a performance measure.

## Observations on the terrain's features relevant to managing and improving assessment systems

Having identified these inputs and outputs, it is now possible to examine each of these more closely (see Fig. 6) to consider what should be thought out anew in assessment systems' processes generally and especially related to the creation of process data specifically.

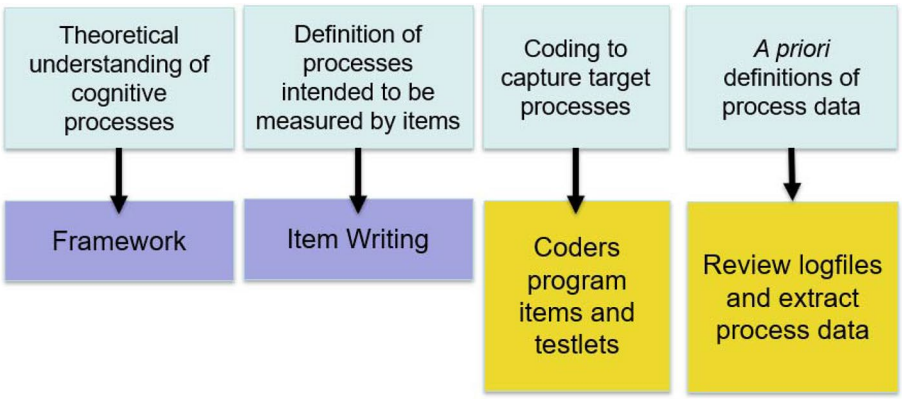


**Fig. 6** Inputs shaping process data. Graphical representation of inputs that can shape and structure process data. Author's own work

Beginning with the input "Theoretical understandings of cognitive processes" which should inform the framework, it should be self-evident that there is a need to identify and define ahead of time which cognitive (and any non-cognitive) processes are important to focus on to ensure that one actually collects data on them. This requires some deep thinking about what exactly is being, or should be, measured. It also runs the possible risk of fracturing domains, like math if process data begin to reveal that there are different cognitive processes that are used in, say, geometric thinking versus algebraic thinking.

The identification of the input "Definition of the processes that are intended to be measured by items" asserts that such definitions should become a standard part of item writing. Traditionally, this has not been part of the practice of item writing and such a change in standard item writing practice may not be as easy as it might seem because there is not yet sufficient knowledge about which processes matter or how to parse cognitive processes into measurable parts. Moreover, item writers are not typically trained to think about a domain's cognitive processes but rather are subject matter experts. There will need to be a feedback loop of some sort so that as assessments evolve there is an iteratively tighter focus on *which* processes to measure—informed both by 'what we want to know' and 'what we can actually measure.' Furthermore, care will need to be taken to validate all information collected on cognitive processes. For example, if some individuals can do items in their head, then the target processes would not be observable or measurable for those individuals, thus skewing the results.

The input "Coding to capture target processes" needs to be done in a way that is standardized not only so that processes are defined consistently, but also so that those who analyze process data on the back end know what they are getting. Currently, process data have often been coded with variable names such as "acer_event_1", "pde_1", etc. which are as unhelpful to analysts as they are obscurely labeled. Exactly who should standardize labeling and coding is an open question, but it needs to be thought out along with (a) a standard operating procedure for transmitting the coding decisions to future analysts and (b) the choice of who should decide which events should be logged, stored,

and exported in the first place—the item writers, the item coders, the system programmers, the assessment management team, some other party?

Finally, the input "A priori definitions of process data" calls for the creation of categories for process data based on some sort of theoretical framework rather than a posteriori definitions determined by data mining, fishing, or best guesses as to how to partition logfile data (e.g., number of uses of the cancel button, number of help menu hits, time on task—as the example of the NAEP math item showed, seemingly logical findings may not have any real meaning). This part of the creation of process data also needs to be done in the way that can become standardized, again both for consistency across assessments and so that those who analyze process data on the back end know what they are getting. It also will require a feedback loop so that—like with the input "definition of processes that are intended to be measured"—there is an iteratively tighter focus on *which* processes to measure or categories to capture.

Examining the process data that can be collected—the "outputs" from an assessment system for lack of better terminology (see Fig. 7)—starts with paradata. The output identified as "Paradata for monitoring assessment integrity" is already being collected and used in PIAAC, which collects data not only on the beginning and end times of interviews and assessments, but also GPS or location data for verification purposes.

The output "Process data for forensic purposes" is also already being collected and used in NAEP—as described earlier—to validate and ascertain item reliability.

The output "Process data for research" holds the most promise of all uses if the field can understand and widen its understanding of thinking itself, for example, by identifying patterns of thought related to performance outcomes (e.g., correct and incorrect responses). If patterns of thought can be identified, then they can, in theory, be categorized in terms of patterns which offer a greater probability of success in some task, as well as related to individual characteristics (e.g., personality, socio-emotional aptitudes, social/familiar upbringing, and culture).

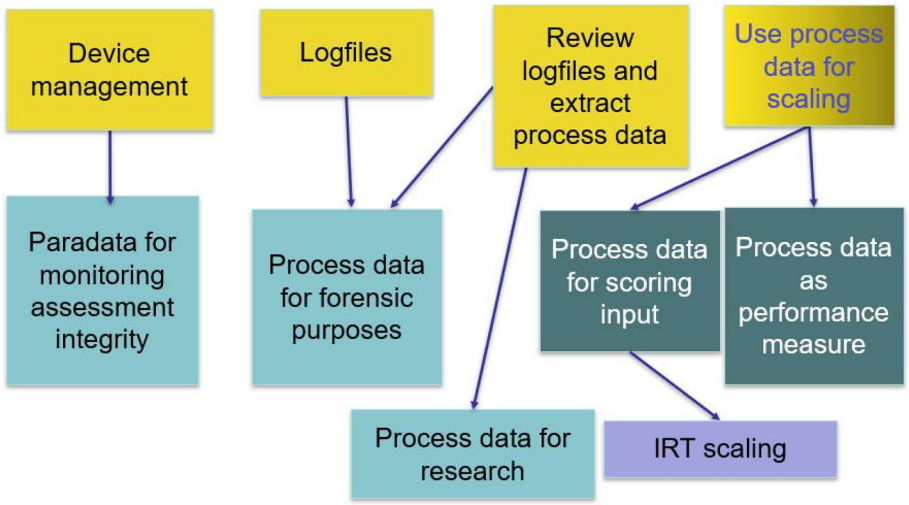


**Fig. 7** Process data outputs. Graphical representation of system outputs relying on process data. Author's own work

This has already been done in some sophisticated though preliminary and exploratory ways. For example, Ido Roll has presented results from research that led to the monograph "Identifying Productive Inquiry in Virtual Labs Using Sequence Mining" (Perez et al. 2017). In this study, about 100 students were given a Direct Current (DC) circuit construction kit, and their strategies for learning during their unstructured activity with this electronics kit were analyzed. Their actions were categorized (a posteriori) into the following types of actions:

Construct, Pause, Test with 1 resistor, Test with 2 resistors, Test with multiple resistors.

Patterns among these actions were compared for students who successfully learned to use the electronics kit to produce working circuits with those who did not demonstrate "productive learning." They found, in short, that patterns with actions in this order: Pause, Test with 2 resistors, Pause, Construct were associated with "productive learners" at higher rates than "unproductive learners" (20 out of 38 for the former, 7 out of 36 for the latter). They also found that Pause, Test with many resistors, Construct, Test with many resistors, Construct was a pattern more common among "unproductive learners" than among "productive learners." The results are reported to show that "a strategic use of pauses so that they become opportunities for reflection and planning is highly associated with productive learning."

While an impressive empirical learning study with concrete processes both identified and put into patterns, the investigation's limited and narrow scope and very contextualized findings suggest the immensity of the task of using process data for research into thinking. Finally, it should be noted that the goal of such research should not be to try to find the "one" best pattern of thought, per se, but to deepen our understanding of thinking.

Turning then to the last two outputs, "process data for scaling" and "process data as a performance measure," it is important to realize that these aspirational uses for process data will require a great deal of thought because they can create some of the greatest hazards. Unless users are told what is expected of them in an assessment, it is challenging to interpret their behavior and indeed potentially unethical to take points away from them for not doing something they did not know they were expected to do. For example, is it ethical to score the same response differently if response time is the only difference between two responses and the respondents did not know that speed mattered? Furthermore, will all countries agree on what certain processes mean and how they can be reported? In PIAAC, anxiety about what could be done with socio-emotional data collected on extroversion and inversion has prompted some Asian countries to demur on administering these items. That said, the potential of moving beyond reporting merely correct outcomes and being able to quantify skill use and practices holds great possibility for improving and transforming assessments and policy-relevant data.

## Summary

This reconnaissance and mapping of the new terrain in assessment systems—in which process data emerge—is in no way meant to be a definitive tour but rather a way to provide some common grounds for future discussions so that those at this conference and in the field can be clearer and more precise from now on in conversations and writing

that refine our ideas, definitions, concepts, and theories about process data, and especially about how process data figures into the development and operationalization of better assessment systems.

The conceptual framework presented here is meant to be broad and flexible enough both for conceptualizing and mapping out further places for process data and for explaining process data to non-researchers. This paper has also pointed out potential pitfalls in various areas and tried to make clear the importance of (1) keeping in mind the underlying constructs (and purposes) of the particular type of process data under discussion and (2) being specific as to (a) which type of process data in the presented schema is under discussion and (b) what goes into that type of process data. However, this paper assuredly has omitted from this framework some important elements and missed critical points. Most obviously, there has been no discussion of the need for a list of inappropriate uses of process data, which will be needed along with standards and guidelines. At a minimum, these uses could include:

- Overgeneralizing from one item to all items, or one process to many processes;
- Concluding that strategies associated with higher performance are the strategies that all students should be taught; and
- Making classroom and formative assessments turn on process data in such a way that students lose unstructured opportunities to try out new ways of thinking and doing.

There are certainly many more inappropriate uses for process data, which this conference may begin to identify. In regard to all the other important and valuable points that have been overlooked in this Keynote address, those hopefully will be covered in comments and commentary during this conference.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s40536-020-00092-z.

**Additional file 1.** Online visualization of "How Time is Spent During Testing" based on process data from NAEP 2017 grade 8 item.

**Acknowledgements**
Ruhan Circi and Fusun Sahin, AIR. The author has obtained permission to acknowledge all those mentioned in the Acknowledgements section.

**Authors' contributions**
SP prepared the entirety of the article without collaboration with or contributions from other authors. The views expressed are those of the author and do not reflect any official policy or position of the Institute of Education Sciences, the U.S. Department of Education, or the U.S. Government. The author read and approved the final manuscript.

**Funding**
No funding was received by the author for the development of this article or for the preparation of the original Keynote Address.

**Availability of data and materials**
Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study. All shown data come from public sources.

**Competing interests**
The author declares that he has no competing interests.

Received: 4 June 2020 Accepted: 23 October 2020
Published online: 04 January 2021

**References**

Bergner, Y., & von Davier, A. (2018). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics., 44*(6), 107699861878470.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of The Royal Society B, Biological Sciences, 205*(1161), 581–598.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., & Stigler, J. (2003). Teaching mathematics in seven countries: Results from the TIMSS 1999 video study. (NCES Report No. 2003-013). Washington, D.C.: National Center for Education Statistics, Institute of Education Sciences, U.S. Dept. of Education.

National Center for Education Statistics. (2020). Process Data From the 2017 NAEP Grade 8 Mathematics Assessment. (NCES Report No. 2020-068). Washington, D.C.: National Center for Education Statistics, Institute of Education Sciences, U.S. Dept. of Education.

Perez, S., Massey-Allard, J., Butler, D., Ives, J., Bonn, D., Yee, N., & Roll, I. (2017). Identifying productive inquiry in virtual labs using sequence mining. In E. André, R. Baker, X. Hu, M. M. T. Rodrigo, & B. du Boulay (Eds.), *Artificial intelligence in education: Proceedings of the 18th international conference, AIED 2017* (pp. 287–298). Springer.

## Publisher's Note