

Process Data Analysis in ILSAs: An Ecological Framework and Literature Review

Denise Reis Costa<sup>1</sup>, Waldir Leoncio Netto<sup>2</sup>

<sup>1</sup> Centre for Educational Measurement (CEMO), University of Oslo

<sup>2</sup> Oslo Center for Biostatistics and Epidemiology (OCBE), University of Oslo

**Author Note**

Correspondence concerning this chapter should be addressed to Denise Reis Costa,

Gaustadalleen 21, 0373 Oslo, Norway. Email: [d.r.costa@cemo.uio.no](mailto:d.r.costa@cemo.uio.no)

### Abstract

Computational advancements in the last couple of decades have brought forth a new era of international large-scale assessments (ILSAs) in which the administration of computer-based tests is becoming the norm. Beyond collecting the correct/incorrect answers for each item, computer-based assessments are also able to collect a range of actions performed by the respondents in the computer testing application during the course of the test administration. Both respondents' actions—starting a unit, clicking a button, spending time until inputting or submitting an answer, and so forth—and their overall behavior—what they do with their keyboards, mice, and even their own eyes—are recorded as a new set of data called “process data”, which are normally time-stamped and can be stored in so-called log files. There may be plenty of useful insight into the respondent's cognitive process, and process data can potentially become a relevant element in the scoring process of an assessment, validate test score interpretations, to name a few possibilities for the analysis of such data. This chapter aims to contribute to the body of knowledge in this area by offering (1) an introduction to process data, what kind of data it contains, its relation to the cognitive process and how it can be organized into a proposed 6-layered ecological framework that facilitates its analysis; (2) a literature review of 37 seminal and state-of-the-art publications produced in the last decade on the topic, which are then analyzed both in their chronological perspective as well as in how they fit into the ecological framework; and (3) a discussion of the potential and limitations of using process data in the assessment framework, including a view of what could be the next steps in the analysis of process data from ILSAs.

*Keywords:* Computer-based assessment; Process data; Log files; Ecological framework; ILSA

### **Introduction**

Over more than two decades of international large-scale assessments (ILSAs), one may notice several structural changes across different cycles of the same assessment. Improvements on test design, changes in the number of evaluated educational systems, and the transition from paper-based to computer-based testing are examples of how an assessment evolves through time. The move toward a more digitalized test administration seems to be a trend of the assessments of the 21<sup>st</sup> century, and it is here to stay.

The Programme for International Student Assessment (PISA), for example, pioneered the implementation of computer-based testing in 2006 (OECD, 2010); since 2015, electronically-delivered assessments are the main mode of administration of PISA tests (OECD, 2017). In 2019, the IEA Trends in International Mathematics and Science Study (TIMSS) began the transition to digital format of assessments (Mullis & Martin, 2017).

With this shift, new forms of data started to emerge from the administration of the test in the computer. In this chapter, we will focus on describing this so-called process data, which represent all the information extracted from and potentially analyzed by a computer platform. Given the facilitated accessibility of such data, researchers of multiple institutions and universities are increasingly overcoming the challenges of the exploration of these complex datasets such that the body of literature with empirical results from the analysis of process data from international large-scale assessments is being published to an increasing degree in the preceding years. The motivation for this chapter comes from noticing the need for evaluating the current state-of-the-art, identifying common practices and possible directions in research and technological developments in the field.

In this work, we give an overview of current developments in the analysis of process data from ILSAs, as well as propose a framework for the analysis given the specificities of these assessments. First, we provide a definition of process data, possible indicators, highlight its significance and propose an ecological framework for its analysis in the context of ILSAs. Second, we summarize empirical research that have used such data to address issues related to test-taking behavior and strategies followed by respondents when answering test items in light of the ecological framework. Finally, we discuss the potential and limits in the analysis of process data as well as take a look at possible developments and research on the field.

### **What are process data?**

To generate public-use files from an assessment, data are processed following several steps from a data management protocol. In PISA 2015, for example, countries that chose the paper-based assessments had to manually entry data from paper forms and booklets to in a specific software,

## PROCESS DATA ANALYSIS IN ILSAs

the Data Management Expert (DME; OECD, 2017). In sequence, various data and validation checks are required at national level as well as under international specifications (for an overview of this process, we recommend a look at chapter 10 of the technical report, OECD, 2017). The data entry of test forms for countries in the computer-based version of the assessment, on the other hand, was automatic and countries had the opportunity to not only collect the students' final answers, but also accurately transcribe those responses and related process data (OECD, 2017).

Differently from authors who delimit process data solely as test-taker actions (Lee & Haberman, 2016), we define process data as any type of information (e.g., response actions or timing) recorded on a computer platform into electronic files. This definition is also in line with the work from Klotzke & Fox (2019) and De Boeck & Scalise (2019) to name a few. In a digital-based assessment, for example, process data are generated from log files, also known as paradata (Kroehne & Goldhammer, 2018). These electronic scripts are time-stamped records of the interactions between the user and the software interface. Eye-tracking movements (Krstić et al., 2018) or digital video recordings of talk and gesture (Maddox, 2017) are also examples of possible sources of collection of process data in a survey.

There is no specific format type for generating process data. Log files, for example, usually have large size and heterogeneously-structured content. Any computer program (e.g., a script embedded on a webpage, a smartphone app) can be configured to produce such files. They are originally programmed to store the various events of software development and monitoring, allowing one to follow the logic of the program, at a fine-grained level, while also making debugging easier (Valdman, 2001). As a by-product of a computer-based administration, software developers are challenged with the trade-off between the amount of information to be saved and the amount of available space on storage devices.

Figures 1 to 3 illustrate three types of process data that can be found in educational assessments. While the first two figures represent screenshots of raw data (which describe in detail all the information recorded from an assessment), the third shows an example of a semi-processed log file (which was created extracting specific features of the raw log files). All files need, however, suitable software to properly digest their structure and produce a readable output.

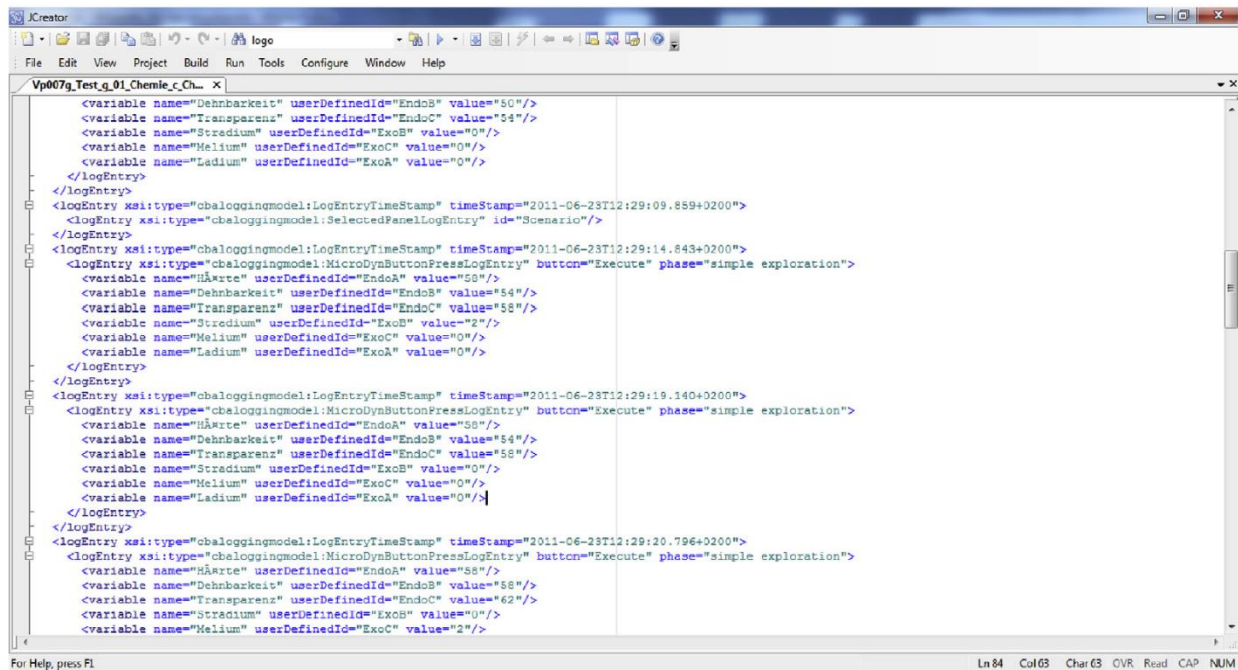
Figure 1 shows an example of raw log file from a complex problem solving (CPS) assessment presented by Greiff, Wüstenberg, & Avvisati, (2015). Using a common extension for generating log files, the Extensible Markup Language (XML) format, Figure 1 displays the interaction of one student with a CPS task. Each chunk of code (inside a "logEntry" tag) includes the timepoint each interaction was executed ("timestamp" command) and the type of interactions (e.g., clicks).

The second image refers to the raw data of eye-tracking movements of one student in a study using PISA items (for more details see Krstić, Šoškić, Ković, & Holmqvist, 2018). Though originally generated in an audio-visual format (i.e., MIDI type), the Intermediate Data Format (IDF) file was transformed to a pure text file which contains information such as head and eye position of the student during the test administration. Differently from log files, process data derived from this

## PROCESS DATA ANALYSIS IN ILSAs

technology record information associated to student eye movement and software pre-fixed features only, not informing specific functionalities of the assessment platform (e.g., clicks, button activated).

In the third screenshot, it is possible to visualize a semi-processed log file in SPSS format with a database structure that is much familiar to many researchers in Education (students in row and variables in columns). Each user activity (e.g., “start item”, “end item”) is recorded as an event. Differently from Figures 1 and 2, this log file illustrates the record of specific events (e.g., clicking the “apply” button) from all students that took the “Climate Control” item in the PISA 2012 assessment. We defined this log file as semi-processed, since it is already organized in a structure that facilitates the statistical analysis, including the selection of variables/features that may be of interest to the researchers. Nonetheless, there is still a necessity for the use of sophisticated procedures and computational resources to analyze data in such format, given their large size and non-trivial interpretation.



```
<variable name="Dehnbarkeit" userDefinedId="EndoB" value="50"/>
<variable name="Transparenz" userDefinedId="EndoC" value="54"/>
<variable name="Stradium" userDefinedId="ExoB" value="0"/>
<variable name="Helium" userDefinedId="ExoC" value="0"/>
<variable name="Ladium" userDefinedId="ExoA" value="0"/>
</logEntry>
</logEntry>
<logEntry xsi:type="cbalogggingmodel:LogEntryTimeStamp" timeStamp="2011-06-23T12:29:09.859+0200">
<logEntry xsi:type="cbalogggingmodel:SelectedPanelLogEntry" id="Scenario"/>
</logEntry>
<logEntry xsi:type="cbalogggingmodel:LogEntryTimeStamp" timeStamp="2011-06-23T12:29:14.843+0200">
<logEntry xsi:type="cbalogggingmodel:MicroDynButtonPressLogEntry" button="Execute" phase="simple exploration">
<variable name="Härte" userDefinedId="EndoA" value="58"/>
<variable name="Dehnbarkeit" userDefinedId="EndoB" value="54"/>
<variable name="Transparenz" userDefinedId="EndoC" value="58"/>
<variable name="Stradium" userDefinedId="ExoB" value="2"/>
<variable name="Helium" userDefinedId="ExoC" value="0"/>
<variable name="Ladium" userDefinedId="ExoA" value="0"/>
</logEntry>
</logEntry>
<logEntry xsi:type="cbalogggingmodel:LogEntryTimeStamp" timeStamp="2011-06-23T12:29:19.140+0200">
<logEntry xsi:type="cbalogggingmodel:MicroDynButtonPressLogEntry" button="Execute" phase="simple exploration">
<variable name="Härte" userDefinedId="EndoA" value="58"/>
<variable name="Dehnbarkeit" userDefinedId="EndoB" value="54"/>
<variable name="Transparenz" userDefinedId="EndoC" value="58"/>
<variable name="Stradium" userDefinedId="ExoB" value="0"/>
<variable name="Helium" userDefinedId="ExoC" value="0"/>
<variable name="Ladium" userDefinedId="ExoA" value="0"/>
</logEntry>
</logEntry>
<logEntry xsi:type="cbalogggingmodel:LogEntryTimeStamp" timeStamp="2011-06-23T12:29:20.796+0200">
<logEntry xsi:type="cbalogggingmodel:MicroDynButtonPressLogEntry" button="Execute" phase="simple exploration">
<variable name="Härte" userDefinedId="EndoA" value="58"/>
<variable name="Dehnbarkeit" userDefinedId="EndoB" value="58"/>
<variable name="Transparenz" userDefinedId="EndoC" value="62"/>
<variable name="Stradium" userDefinedId="ExoB" value="0"/>
<variable name="Helium" userDefinedId="ExoC" value="2"/>
```

Figure 1: Screenshot of a raw .xml file from a computer-based assessment (from Greiff, Wüstenberg, & Avvisati, 2015).

# PROCESS DATA ANALYSIS IN ILSAs

```

1 ## [iView]
2 ## Converted from: BATMAN-eye data.idf
3 ## Date: 12.05.2017 15:50:56
4 ## Version: IDF Converter 3.0.20
5 ## IDF Version: 9
6 ## Sample Rate: 60
7 ## Separator Type: Msg
8 ## Trial Count: 1
9 ## Uses Plane File: False
10 ## Number of Samples: 52718
11 ## Reversed: none
12 ## [Run]
13 ## Subject: BATMAN
14 ## Description: SM VIII1
15 ## [Calibration]
16 ## Calibration Area: 1920 1080
17 ## Calibration Point 0: Position(960;540)
18 ## Calibration Point 1: Position(480;10)
19 ## Calibration Point 2: Position(1900;270)
20 ## Calibration Point 3: Position(1440;1069)
21 ## Calibration Point 4: Position(19;810)
22 ## [Geometry]
23 ## Stimulus Dimension [mm]: 345 195
24 ## Head Distance [mm]: 640
25 ## [Hardware Setup]
26 ## System ID: RADISA
27 ## Operating System : 6.2
28 ## iView X Version: 2.11.71
29 ## [Filter Settings]
30 ## Heuristic: False
31 ## Heuristic Stage: 0
32 ## Bilateral: True
33 ## Gaze Cursor Filter: True
34 ## Saccade Length [px]: 80
35 ## Filter Depth [ms]: 20
36 ## Format: LEFT, RIGHT, RAW, DIAMETER, CR, POR, QUALITY, PLANE, HEADPOSITION, HEADROTATION, EYEPOSITION, GAZEVECTOR, MSG, FRAMECOUNTER
37 ##
38 Time Type Trial L Raw X [px] L Raw Y [px] R Raw X [px] R Raw Y [px] L Dia X [px] L Dia Y [px] L Mapped Diameter [mm] R Dia X [px] R Dia Y [px]
39 Validity R Validity Pupil Confidence L Plane R Plane H POS X [mm] H POS Y [mm] H ROT X [°] H ROT Y [°] H ROT Z [°]
40 13180604585 SMP 1 547.82 599.21 792.96 595.82 16.87 16.87 16.69 16.69 4.64 546.17 602.11 551.47 602.10 789.63 599.15 794.86
41 13180613957 MSG 1 # UTC: 1494604256 127
42 13180625634 SMP 1 547.97 599.29 792.94 595.90 16.68 16.68 16.64 16.64 4.64 546.11 602.19 551.45 602.16 789.59 599.18 794.82
43 13180637558 SMP 1 547.75 599.27 792.79 595.85 16.91 16.91 16.55 16.55 4.63 546.07 602.20 551.35 602.19 789.52 599.16 794.77
44 13180645902 SMP 1 547.62 599.25 792.77 595.84 16.91 16.91 16.42 16.42 4.62 546.00 602.15 551.19 602.11 789.43 599.08 794.68
45 13180678334 SMP 1 547.57 599.12 792.78 595.60 16.89 16.89 16.71 16.71 4.62 545.97 602.11 551.17 602.10 789.37 599.02 794.62
46 13180708219 SMP 1 547.45 599.11 792.60 595.58 16.94 16.94 16.73 16.73 4.62 545.92 602.06 551.08 602.04 789.29 598.97 794.48
47 13180728964 SMP 1 547.36 599.21 792.61 595.76 16.71 16.71 16.74 16.74 4.63 545.89 602.11 551.03 602.12 789.14 599.10 794.43
48 13180761218 SMP 1 547.36 599.30 792.62 595.94 17.02 17.02 16.90 16.90 4.64 545.77 602.20 551.02 602.23 789.14 599.25 794.39
49 13180773812 SMP 1 547.37 599.29 792.49 596.01 17.12 17.12 16.81 16.81 4.64 545.72 602.28 550.97 602.29 789.12 599.27 794.35
50 13180790506 SMP 1 547.36 599.36 792.43 595.94 17.14 17.14 16.76 16.76 4.64 545.66 602.32 550.92 602.29 789.08 599.27 794.19
51 13180804466 SMP 1 547.18 599.33 792.38 595.94 17.10 17.10 16.77 16.77 4.64 545.58 602.28 550.86 602.25 789.03 599.20 794.13

```

Figure 2: Screenshot of a raw .idf file from Krstić, Šoškić, Ković, & Holmqvist (2018) eye-tracking study (data extracted from <https://osf.io/xjd5r/>)

Case	cnt	school	SIDStd	event	time	event_number	event_type	tsp_setting	central_setting	bottom_setting	temp_value	humid_value	diag_state	var	var	var	var	var	var	var	
1	ARE	0000189	04852	START_ITEM	1298.1000	1	NULL	NULL	NULL	NULL	NULL	NULL	NULL								
2	ARE	0000189	04852	ACER_EVENT	1291.9000	2	reset	0	0	0	25	25	NULL								
3	ARE	0000189	04852	ACER_EVENT	1338.4000	3	apply	1	1	1	27	28	NULL								
4	ARE	0000189	04852	ACER_EVENT	1346.8000	4	apply	1	1	2	29	33	NULL								
5	ARE	0000189	04852	ACER_EVENT	1352.0000	5	apply	1	2	2	31	36	NULL								
6	ARE	0000189	04852	ACER_EVENT	1354.5000	6	apply	2	2	2	35	36	NULL								
7	ARE	0000189	04852	ACER_EVENT	1361.1000	7	apply	2	1	1	36	36	NULL								
8	ARE	0000189	04852	ACER_EVENT	1361.1000	8	reset	0	0	0	25	25	NULL								
9	ARE	0000189	04852	ACER_EVENT	1375.3000	9	Diagram	NULL	NULL	NULL	NULL	NULL	000000								
10	ARE	0000189	04852	ACER_EVENT	1376.2000	10	Diagram	NULL	NULL	NULL	NULL	NULL	000000								
11	ARE	0000189	04852	ACER_EVENT	1400.1000	11	Diagram	NULL	NULL	NULL	NULL	NULL	000000								
12	ARE	0000189	04852	ACER_EVENT	1402.1000	12	Diagram	NULL	NULL	NULL	NULL	NULL	000001								
13	ARE	0000189	04852	ACER_EVENT	1406.8000	13	Diagram	NULL	NULL	NULL	NULL	NULL	000001								
14	ARE	0000189	04852	ACER_EVENT	1408.4000	14	Diagram	NULL	NULL	NULL	NULL	NULL	000011								
15	ARE	0000189	04852	ACER_EVENT	1410.2000	15	Diagram	NULL	NULL	NULL	NULL	NULL	000101								
16	ARE	0000189	04852	ACER_EVENT	1410.6000	16	Diagram	NULL	NULL	NULL	NULL	NULL	000101								
17	ARE	0000189	04852	END_ITEM	1416.1000	17	NULL	NULL	NULL	NULL	NULL	NULL	000101								
18	ARE	0000189	04861	START_ITEM	1309.1000	1	NULL	NULL	NULL	NULL	NULL	NULL	NULL								
19	ARE	0000189	04861	ACER_EVENT	1336.8000	2	Diagram	NULL	NULL	NULL	NULL	NULL	000000								
20	ARE	0000189	04861	ACER_EVENT	1336.9000	3	Diagram	NULL	NULL	NULL	NULL	NULL	000000								
21	ARE	0000189	04861	ACER_EVENT	1338.9000	4	Diagram	NULL	NULL	NULL	NULL	NULL	000000								
22	ARE	0000189	04861	ACER_EVENT	1346.9000	5	Diagram	NULL	NULL	NULL	NULL	NULL	000000								
23	ARE	0000189	04861	ACER_EVENT	1358.7000	6	Diagram	NULL	NULL	NULL	NULL	NULL	000000								
24	ARE	0000189	04861	ACER_EVENT	1360.1000	7	Diagram	NULL	NULL	NULL	NULL	NULL	000000								
25	ARE	0000189	04861	ACER_EVENT	1361.0000	8	Diagram	NULL	NULL	NULL	NULL	NULL	000001								
26	ARE	0000189	04861	ACER_EVENT	1362.0000	9	Diagram	NULL	NULL	NULL	NULL	NULL	000001								
27	ARE	0000189	04861	ACER_EVENT	1366.5000	10	Diagram	NULL	NULL	NULL	NULL	NULL	000001								
28	ARE	0000189	04861	ACER_EVENT	1367.2000	11	Diagram	NULL	NULL	NULL	NULL	NULL	000001								
29	ARE	0000189	04861	END_ITEM	1378.9000	12	NULL	NULL	NULL	NULL	NULL	NULL	000101								
30	ARE	0000189	04843	START_ITEM	191.1000	1	NULL	NULL	NULL	NULL	NULL	NULL	000101								
31	ARE	0000189	04843	ACER_EVENT	226.6000	2	reset	0	0	0	25	25	NULL								
32	ARE	0000189	04843	ACER_EVENT	248.2000	3	apply	-2	2	2	21	31	NULL								
33	ARE	0000189	04843	ACER_EVENT	243.0000	4	apply	-2	2	-2	17	29	NULL								
34	ARE	0000189	04843	ACER_EVENT	246.1000	5	apply	-2	-2	-2	13	23	NULL								
35	ARE	0000189	04843	ACER_EVENT	248.3000	6	apply	2	-2	-2	17	17	NULL								
36	ARE	0000189	04843	ACER_EVENT	250.6000	7	apply	2	2	-2	21	15	NULL								
37	ARE	0000189	04843	ACER_EVENT	252.6000	8	apply	2	2	2	25	21	NULL								
38	ARE	0000189	04843	ACER_EVENT	260.7000	9	apply	1	-1	1	27	22	NULL								
39	ARE	0000189	04843	ACER_EVENT	266.7000	10	apply	-1	1	-1	25	21	NULL								
40	ARE	0000189	04843	ACER_EVENT	270.6000	11	apply	-1	-1	-1	23	18	NULL								
41	ARE	0000189	04843	ACER_EVENT	277.8000	12	apply	1	1	1	25	21	NULL								
42	ARE	0000189	04843	ACER_EVENT	280.9000	13	reset	0	0	0	25	25	NULL								
43	ARE	0000189	04843	ACER_EVENT	283.8000	14	Diagram	NULL	NULL	NULL	NULL	NULL	000000								

Figure 3: Screenshot of a semi-processed .sav file from PISA 2012 Climate control item (data extracted from <https://www.oecd.org/pisa/pisaproducts/database-cbapisa2012.htm>)

### **What kind of information can we get from process data?**

During test administration, raw log files collect in a group of lines (e.g., a chunk of code from the XML file) data such as user interaction, function call, etc. All data records are gradually appended to the file, never deleting or changing stored information (Valdman, 2001). From this tangle of information, two types of operational variables are especially important for the analysis of process data in ILSAs: time and respondent's actions.

Each information reported in a log file is listed in chronological order. From the timing information, one can calculate the total time respondents spent an item before giving their final answer (also known as "time on task" or "response time"), the time the respondent spent since their first interaction with the test platform, and so on. It is usually up to the test programmer or due to software constraints to define the unit of measurement of the time variables (e.g., seconds, milliseconds).

We define as respondent's action any interaction (single or multiple) between users and the test platform. From a simple mouse click to a sophisticated interaction in a simulated-based task, respondent's actions can reveal the actual respondent behavior that led to the performance outcome (Herde et al., 2016). The type and unit of measurement of the respondent's actions vary according to the specificities of the test items or what the test developers consider important or useful to collect. For instance, in the Climate control item from Figure 3, not all interactions between a student and the test environment are recorded in the log file (Chen et al., 2019). From the available information, one can define as respondent's actions the following variables: "top\_setting", "central\_setting", "bottom\_setting", and "diag\_state". Except for these four variables related to respondent's actions and the variable time, the remaining variables were created by the programmer and they do not represent students interactions with the test platform (e.g., "cnt", "schoolid"). Variables "temp\_value" and "humid\_value", for example, represent pre-determined output values that were displayed on the screen, but were solely dependent on the input variables associated to respondent's actions (i.e., "top\_setting", "central\_setting", or "bottom\_setting").

Even though several measures can be extracted from respondent's actions and timing information, the ones that reflect respondent's behavior during the assessment with respect to a particular latent process (e.g., test-taking disengagement as the response time below a certain threshold) are called process indicators (Goldhammer & Zehner, 2017). The finer-grained information recorded in the log files, the greater the potential for the investigation of underlying cognitive processes and respondent's strategies, as well as the analysis of construct-irrelevant variation on the test scores, such as test-taking engagement (Goldhammer et al., 2016).

It is also important to note that personal information can be gathered in such files and ethical approval and consent for data collection may be necessary. For the PISA assessment, for example, OECD and the participating countries or economies establish an agreement regarding the public release of the data collected in the assessment in which information such as micro-level data (e.g., student and school identification) is anonymized (OECD, 2017).

### **How are process data related to response process?**

There is an increasing interest in the analysis of process data, since it can provide a non-evasive way to observe how respondents solve the test items. It can also be used to proxy unobservable traits (e.g., test-taking motivation) to better understand the relationship between these attitudes and performance (OECD, 2019).

Even though a well-planned log file can capture fine-grained information about respondent's behavior, there is still a substantial difference between overt behavior and its underlying response processes (Greiff et al., 2015). We agree with Hahnel, Goldhammer, Naumann, & Kröhne (2016) that the operationalization of a test-taking behavior (e.g., navigation) from process data is not a direct measure of respondents' cognitive processes but rather the result of them. Especially in the context of ILSAs, where a large number of educational systems is evaluated on each cycle with respondents coming from different societal contexts, the understanding of the test-taker response process is of fundamental importance to the analysis of process data.

To define response process, we follow Hubley & Zumbo (2017):

*“...one may think broadly of response processes as the mechanisms that underlie what people do, think, or feel when interacting with, and responding to, the item or task and are responsible for generating observed test score variation. This definition expands response processes beyond the cognitive realm to include emotions, motivations, and behaviors. Inclusion of affect and motives allows us to take into account how these may impact the different respondents' interactions with the item(s), test, and testing situation. Our definition also requires one to go beyond the surface content of the actions, thoughts, or emotions expressed by, or observed in, respondents to identify the mechanisms that underlie this content. Finally, we encourage researchers and theorist to develop contextualized and dynamic frameworks that take into account the situational, cultural, or ecological aspects of testing when exploring evidence based on response processes.”*

In this sense, response process from ILSA process data can be seen as more than a collective of mental operations. Thus, evidence-based research can be placed in a wider context, where process indicators and test performance is not only attributed to the respondents or test settings, but also to intertwined factors such as personal traits and social context (e.g., school or educational system).

Based on an evolutionary and adaptive view of human interaction with their environment, Chen & Zumbo (2017) proposed an ecological framework where item responses and test performance is the by-product of the relationship between individuals and their context. It is guided by an abductive explanation for the variation in test performance (Mislevy, 1994; Stone & Zumbo, 2016). In their conceptual model, contextual factors are organized in such a way that the sources of item response or test performance variability can be studied systematically. Expanding this framework to the process data analysis from ILSA is the aim of the next section.



### **An ecological framework for the analysis of process data**

Borrowing the knowledge from ecological systems theory, we propose a conceptual framework for process data analysis using the work from Chen & Zumbo (2017)—who have proposed an ecological framework for item responding and test performance—as an initial reference. Such framework is structured in layers where each piece represents a unit of analysis and allows one not only to better understand how diverse the analysis of process data can be (not only with respect to techniques, but to the scope they reach), but also how these different parts relate to each other and the whole framework.

A graphical representation of our 6-layered framework can be seen in Figure 4. This conceptual model is an adaptation of Chen & Zumbo (2017) proposal, which defines five layers to represent an explanation for variation in testing results: (1) test and test setting characteristics; (2) personal characteristics; (3) classroom and school context; (4) family ecology or other outside-of-school ecology; and (5) characteristics of the educational system and the national state.

Beyond item responses or test performance, we expand Chen & Zumbo (2017)'s framework to incorporate operational variables that can be collected in process data from ILSA and may provide additional information to explain test score variability (i.e., response time, respondent's actions). We have also distinguished item and test characteristics in different layers due to their specificities for data management and analysis in the context of process data.

The first layer of the ecological model relates to intrinsic characteristics of a test item. It encompasses features, content, format, and psychometric properties (e.g., difficulty and discrimination) of an item. For example, released log file data from PISA 2012 (see Figure 3) and from the Programme for the International Assessment of Adult Competencies (PIAAC) were delivered by item. Someone interested in exploring such files must have an understanding of the item characteristics to have a better understanding of the variables (e.g., the definition of “top\_setting” in the Climate Control item) presented in the files.

The following layer relates to the test characteristics (including test environment/setting) as a whole or to a specific group of items where there are nested and crossed dependencies within test-taker data (e.g., a testlet, one booklet/test form, one item block or a unit from the test design). Test objectives, duration, contents, and the psychometric properties play an important role in the observed behaviors extracted from the log files and test performance.

## PROCESS DATA ANALYSIS IN ILSAs

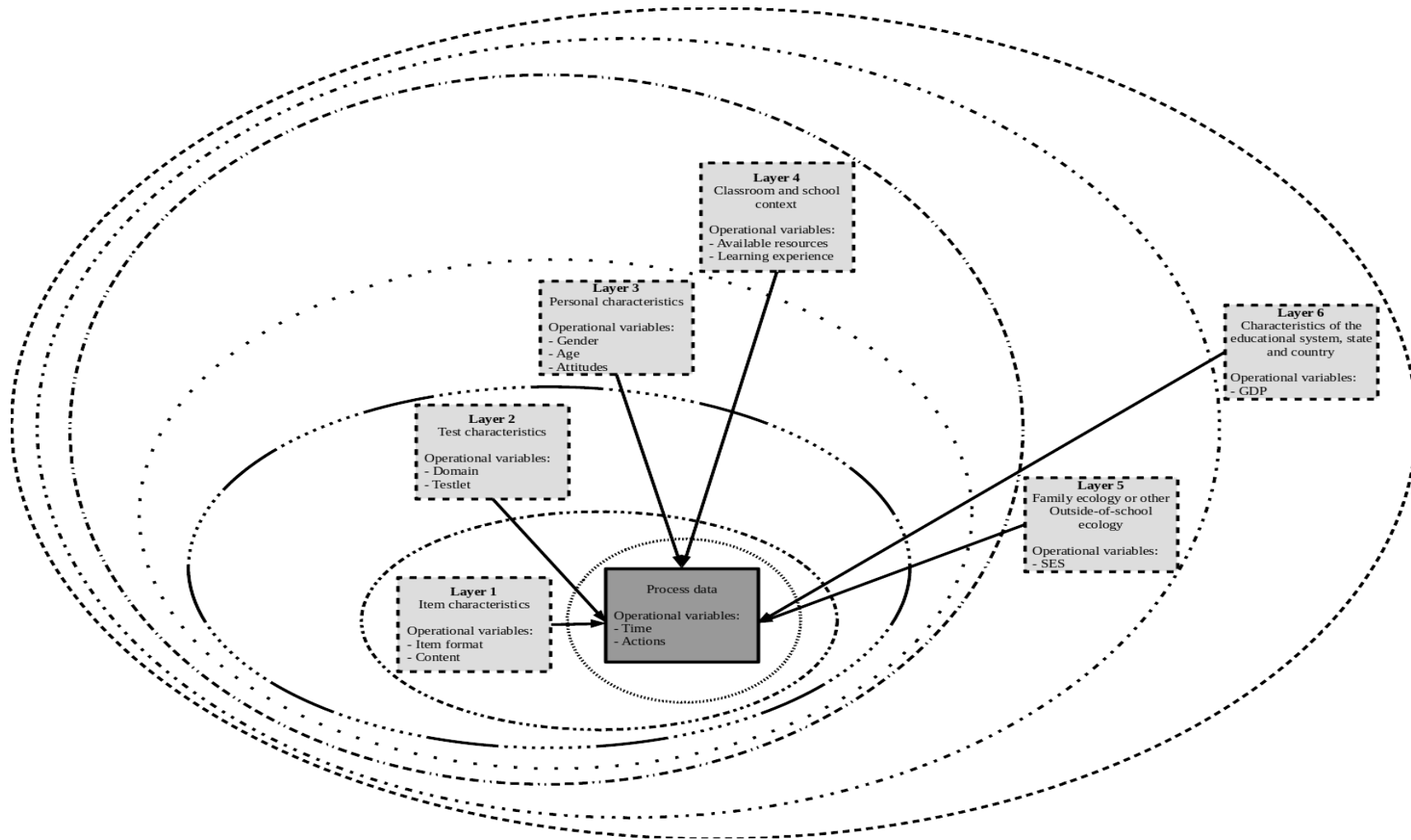


Figure 4: An ecological framework for the analysis of process data in ILSAs with examples of possible operational variables from each layer. An adaptation of the Chen & Zumbo (2017) framework for item responding and test performance.

## PROCESS DATA ANALYSIS IN ILSAs

The third layer is linked to test taker characteristics such as gender identity, age group, and psychosocial traits. Together with the first two layers, personal characteristics and test/item features have the strongest connections between the observed behaviors extracted on the log files and test performance. To a lesser extent, however, other contextualized factors such as school environment (layer 4), family background, out of school experiences (layer 5), as well as the broader context within geographical region and educational system at a national context (layer 6) may also influence how respondents answer to test items and how it is translated in the process data. For instance, students from the same school may share some common strategies when answering computer-based items due to their learning experiences and available resources at school. Extracting new information from process data may also help shed light on the relationship between skills in the population and a measure of economic prosperity (e.g., per capita GDP) as described by van Damme, (2014).

Each layer of this ecological framework can be seen as a unit of analysis. Their representation as a stacked Venn diagram with common centers illustrates the multilevel approach that is intrinsic to the analysis of ILSA datasets (i.e., students nested in schools, schools nested in educational systems). This model also provides a flexible framework for the development of contextual models to explain test results in the sense that one can conduct their analysis using adjacent or non-adjacent layers (e.g., item and personal characteristics – layers 1 and 3) without the necessity to accommodate all layers.

### **How could the proposed ecological framework help the analysis of process data from ILSAs?**

An understanding of the ecological framework can contribute to improve several stages of the data analysis from pre-processing (i.e., the research phase when researchers define their research questions and elaborate a data management plan to extract their operational variables from the raw/semi-processed log-file data) to interpreting and applying of the study's findings. Figure 5 presents the three basic steps that a researcher may encounter when analyzing process data from ILSAs.

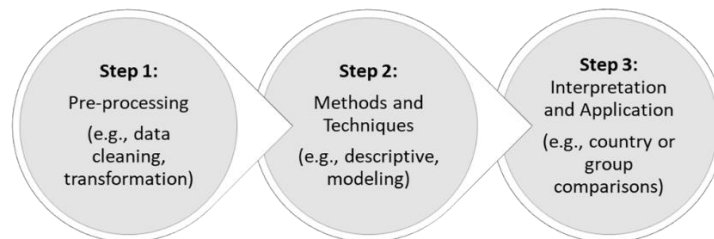


Figure 5: General steps for the analysis of process data

The first step of the analysis of process data is usually a complicated and demanding task. It may include data cleaning, transformation or grouping of operational variables, feature engineering, and so on. Focusing on the research questions at hand, the data pre-processing will usually start

## PROCESS DATA ANALYSIS IN ILSAs

by looking at item-level characteristics (layer 1) but it can also involve more than one layer of the ecological framework. The lower the layer is located in the framework, finer grained information and more data to process. Available tools such as the OECD LogDataAnalyzer (OECD, 2013) and the R package LOGAN (Reis Costa & Leoncio, 2019) have been emerging in the field to help researchers in this stage of analysis.

After data management, myriad approaches and techniques can be used for the analysis of process data and most of them have been already applied successfully in the educational scenario (e.g., psychometric models, data mining techniques). However, the ecological model can still be useful by offering as a visual mapping for the investigation of contextual factors in process data and contributing to its interpretations through the framework's layers (e.g., item/test level or across relevant population's subgroups).

In this study, we present a literature review of empirical studies on process data analysis from ILSAs by the framework's layers. We link the findings to the ecological model to illustrate the usefulness of an understanding of the conceptual framework's layers to connect different analyses strategies and applications.

### Method

#### **What do we know about the analysis of process data from ILSAs?**

In this section, we give a short summary of empirical studies concerning the analysis of process data in ILSAs. First, we describe the criteria for selecting the journal papers, book chapters and working papers analyzed in this review. Then, we present a timeline with key moments in the last decade when the most well-known international surveys started to administer computer-based assessments. In the same fashion, we pinpoint how many studies with empirical results from the analysis of process data from ILSA are situated. Lastly, we categorize the studies following the ecological framework layers. We believe that such analysis will help the reader have a big picture of such studies and understand the level of analysis, associated process indicators, as well as the methods and techniques most used in the field.

#### **Which studies with process data from ILSAs are analyzed here?**

This is a non-exhaustive review that was carried out to address the following question: “What (statistical) approaches and strategies for the analysis of process data from international large-scale assessments are documented in the scientific literature?” For this purpose, we searched journal papers, book chapters and working papers available in the Scopus database between November and December 2019 and conducted a snowballing review to identify additional works. We also used Google Scholar to collect information regarding the number of citations of each paper included in this review.

We identify 21 search terms in which we expected to find all the relevant literature (Table 1). In the search queries, the terms were only connected with the “AND” command and were related to

## PROCESS DATA ANALYSIS IN ILSAs

type of data (e.g., process data, log-file data, computer-generated data), assessment (e.g., PISA, PIAAC, international survey), and analysis (e.g., data analysis, psychometric model, data mining).

Table 1: Search terms used in our literature review

Type	Expressions
Dataset	“Log file”, “Log-file data”, “Paradata”, “Process data”, “Computer generated data”, “Computer-based data”, “Computer-assisted data”
Assessment	“International large-scale assessment”, “International survey”, “International comparative studies”, “International assessment”, “International evaluation”, “PISA”, “TIMSS”, “PIAAC”
Analysis	“Data analysis”, “Data analytics”, “Psychometric”, “Data mining”, “Measurement model”, “Learning analytics”

As inclusion criteria, we analyzed studies published in the English language and that present quantitative analyses with a clear description of the analyzed data and methods for the evaluation of process data from ILSAs. All references were imported in the Rayyan application (Ouzzani, Hammady, Fedorowicz, & Elmagarmid, 2016), which helped with the elimination of duplicate studies that were found in the database searches. After the exclusion of the duplicates and articles that were untraceable, the studies were successively screened based on their relevance to the current review.

## Results

### When did the first studies start to be published?

Table 2 summarizes the 37 studies included in this review. Almost 90% of them were published as journal papers; only two book chapters (He & von Davier, 2015; Ramalingam & Adams, 2018) and two working papers (Goldhammer et al., 2016; He et al., 2019) were accessible in the moment of this literature review.

Table 2

Figure 6 shows the growth of electronically-delivered assessments in ILSAs as well as the rapid increase of published studies in the last decade. From our findings, the first publishing of an empirical study of process data from ILSAs in a scientific journal is dated from 2014 with results from the reading and problem-solving domains from the first round of the PIAAC survey. The study of Goldhammer et al. (2014) is also the one in our review that received the most citations on Google Scholar, 145. From then on, studies using PISA data started to grow in the literature and became prevalent (around 60% of the studies are regarded PISA studies) in the recent years.

Besides the use of OECD datasets, this review included one study using data from the Assessment and Teaching of 21st Century Skills project (ATC21s) performed by Vista, Care, & Awwal (2017) and another involving an international language assessment (with no explicit information about the name of the assessment) analyzed by Lee & Haberman (2016). Although eTIMSS only started

## PROCESS DATA ANALYSIS IN ILSAs

collecting data for the main study in 2019, no studies using possible field trial/testing data were found.

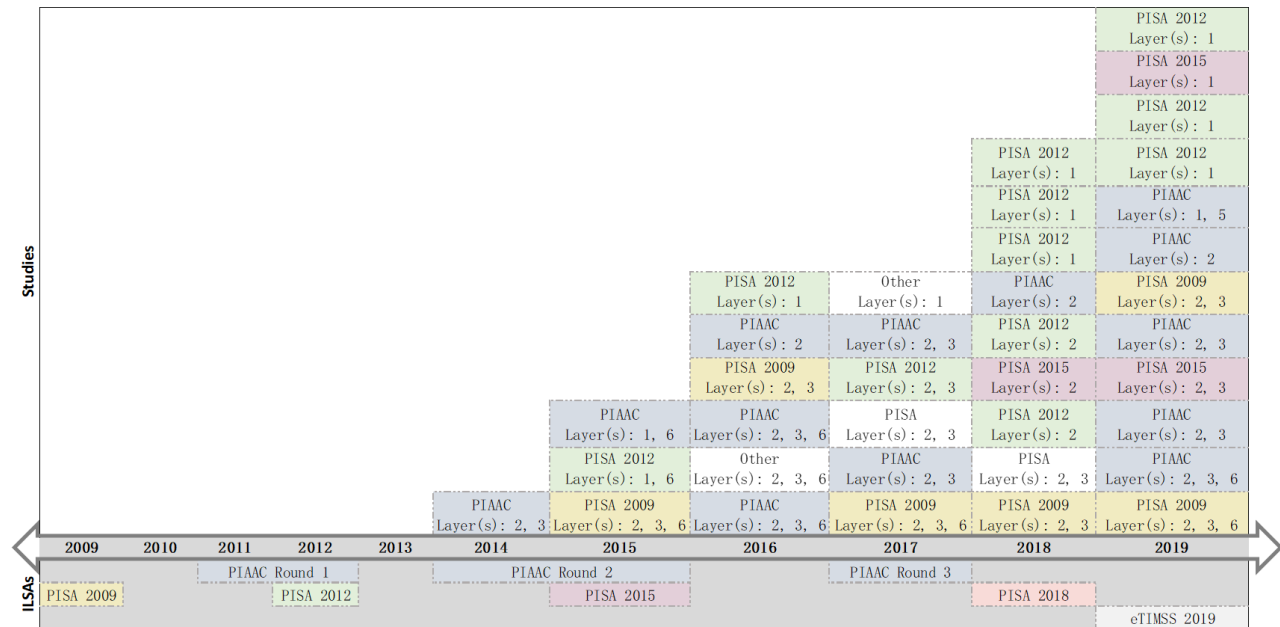


Figure 6: Timeline with information of the year of administration of selected international surveys, as well as the respective datasets used in the scientific publications included in this review ordered by the layers of the ecological framework.

Most of the publications from this study analyzed process data from a group or all items in an assessment (layer 2 from the ecological framework). Problem solving is the domain with the most research on process data, followed by reading. Any studies from this review included the evaluation of process data from the science domain, though there was some research under course by the time of this review (Teig, 2019).

### What approaches and strategies were used in the analyses?

Even though the timing variable is recognized as the main focus of many studies on the analysis of process data from the assessment of cognitive abilities (Greiff et al., 2015), the majority of the quantitative studies included in this review have respondent's actions as key operational variable as well. To get a glimpse of the methodologies and approaches to analyze these behavioral indicators, Figure 7 shows the total number of studies classified by the layers of the ecological framework.

## PROCESS DATA ANALYSIS IN ILSAs

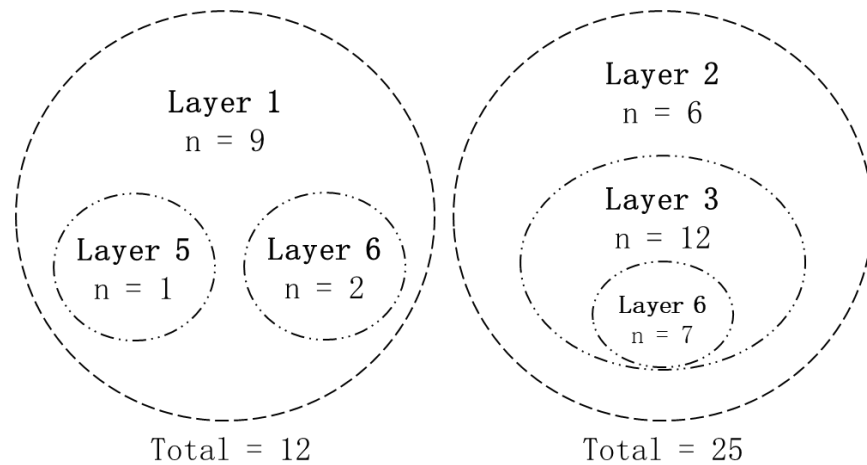


Figure 7: Total number of published works from our literature review grouped by the ecological framework's layers.

None of the 37 papers analyzed in this review focused on layer 4 (“Classroom and school context”). We counted that 12 papers focused on the analysis process data from a specific item (layer 1), with two of them including country-level comparisons (layer 6) and one describing outside-of-school variables (layer 5). When analyzing a group of items or by assessment domain (layer 2), the majority of studies took into consideration the respondent's characteristics (layer 3).

The following sections present the approaches used by the authors when working with process data from each unit of analysis. Although many methods can be applied across studies in different layers of the framework, this exercise aims to provide a quick look on examples of analytical strategies based on the granularity of the data.

### Item-level analysis (layer 1)

Nine out of the 37 studies focused on the analysis of a single problem-solving item from ILSA. Except for the study of Vista et al. (2017), which studied data from the ATC21S project, all journal papers in this layer analyzed data from PISA assessments, with four of them (Chen et al., 2019; Han et al., 2019; Pejic & Molcer, 2016; Xu et al., 2018) exploring respondent's actions in the log file from the same item (Climate Control, Figure 3).

The key feature of these studies is the exploration of log file data using data mining techniques (e.g., naïve Bayes classifier (Pejic & Molcer, 2016), random forest algorithm (Han et al., 2019; Qiao & Jiao, 2018), exploratory network analysis (Vista et al., 2017)); identification of (latent) groups with differential problem solving strategies via a modified multilevel mixture item response theory (IRT) modeling (Liu et al., 2018), latent class analysis (Xu et al., 2018), cluster analysis (Ren et al., 2019); prediction of duration and final outcome via event history analysis (Chen et al., 2019); or investigating the relationship of task performance and observed variables from log-file data through confirmatory factor analysis (De Boeck & Scalise, 2019).

## PROCESS DATA ANALYSIS IN ILSAs

Even though most of the studies in this category analyzed data from more than one participating country in the assessments, studies in this layer focus on the analysis of item characteristics and did not perform any group-level analysis (e.g., cross-country comparisons).

### **Item-level analysis and group-level characteristics (layers 1, 5; layers 1, 6)**

Studies in these layers not only dive in the features of a problem-solving item, but also discuss results in an aggregate format (e.g., country-level analysis).

In the “layer 1, 5” (item and out-of-school variables analysis) category is the work of Liao et al. (2019), which aimed to “provide information for improving competences in adult education for targeted groups”. Specifically, the authors identified action sequences in a problem-solving item via application of natural language processing and text-mining technique (n-gram) using employment-related background variables from the PIAAC study.

Two works were classified in the “layer 1, 6” (item and country-level analysis). While Greiff et al. (2015) discussed the application of an action strategy (VOTAT: vary one thing at a time) in the PISA Climate Control item (Figure 3) and its relation with performance, He & von Davier (2015) explored and investigated how sequences of actions derived from n-grams were related to item-performance in a PIAAC problem-solving task. Both papers conducted an analysis of the test-taker’s strategies by a selection of countries.

### **Group of items/Test-level analysis (layer 2)**

Studies from this categorization represents several assessment domains (e.g., reading, mathematics, problem solving), as well as an exploration of contextual questionnaires from PISA (Kroehne & Goldhammer, 2018). In this layer, process data from ILSAs are analyzed with respect to either a group of items or all the items in a specific domain.

By using response times of 10 math items from PISA 2012, Zhan et al. (2018) evaluated the improvement of model parameter estimates in a cognitive diagnosis approach. Vörös & Rouet (2016), in turn, conducted a logistic regression analysis to study the relationship of performance and observed variables from log-file data (i.e., response time and logged action count) from a selection of tasks from the PIAAC problem-solving domain.

Three works in this layer emphasized the issue of the validity in educational assessment settings using process data. Here, Ramalingam & Adams (2018) investigated how students’ navigation in reading assessment from PISA 2012 could improve the validity and reliability in an IRT framework. Engelhardt & Goldhammer (2019), on the other hand, studied the validation of test scores interpretations using processing times in a structural equation modeling approach with literacy items in PIAAC. With a small-scale study, Maddox et al. (2018) also tackled the validity argument using process data from eye-tracking observations with PIAAC items.



## PROCESS DATA ANALYSIS IN ILSAs

Kroehne & Goldhammer (2018), in turn, analyzed process data through finite-state machines approach (FSM) for questionnaire items besides proposing a framework to classify the information provided in log-files into states and store them as log events.

### **Group of items/test-level analysis and personal characteristics (layers 2, 3)**

A substantial number of studies in this category (5 out of 12) analyzed data from the reading domain in a PISA assessment. Analysis of process data in this category varied between descriptive analyses, latent regressions, IRT, GLMM, and Bayesian covariance structure models.

In the work of Hu et al. (2017), two types of PISA items (analytical and interactive problems) were used to investigate information-process strategies using eye-tracking data from high and low performing groups of students. A descriptive statistical analysis and *t*-test were conducted to evaluate the eye movement differences between the two groups of students for each type of item. Using heat maps, Krstić et al. (2018) also explored the similarities and differences in eye movement patterns between students with high and low scores on PISA reading items. Maddox (2017), in turn, described an exploratory approach via video-ethnographic observations that allows the identification of examples of respondent's fatigue and observing disengagement in a talk and gesture study using PIAAC items.

Structural equation modeling (i.e., latent regression and mediation models) was the statistical approach used by Hahnel et al. (2016) to investigate individual differences in students' skills in comprehending digital text by their navigation behavior and various underlying skills (e.g, basic computer skills, evaluation of online information).

Using response times from ILSAs in a joint modeling framework with response accuracy, Kroehne et al. (2019) evaluated the invariance of response processes regarding the assessment mode (i.e., computer-based vs. paper-based) and the respondent's gender. With an online publication appearing in 2019, the work of Ulitzsch et al. (2020a), in turn, aimed to develop a framework that incorporates respondent's nonresponse behavior to gain a deeper understanding of the processes underlying item omissions in large-scale assessments. Later, the same authors expanded the approach to incorporate test-taking engagement behavior (Ulitzsch et al., 2020b).

In the same category, four papers used an explanatory IRT approach via generalized linear mixed modelling (GLMM) for the analysis of process data with ILSA items. The work of Goldhammer et al. (2014), for example, discusses the application of GLMM to investigate the role of time on task and item and person characteristics (e.g., relative easiness, cognitive operations) on performance for two PIAAC domains: reading and problem solving. Hahnel et al. (2018), in turn, used respondent's actions (i.e., navigation through links from a search engine result page) to analyze how individual differences in reading skills on word, sentence, and text level affect students' ability to evaluate online information. In 2017, Hahnel and colleagues used the same statistical approach for regressions of the dichotomous digital reading scores on several predictors at the student's level (e.g., memory updating, linear reading) and item level (e.g., number of target

## PROCESS DATA ANALYSIS IN ILSAs

and irrelevant nodes). Goldhammer et al. (2017), on the other hand, used person-level variables such as gender, age group, and educational attainment to explain differences in test-taking engagement for a sample of Canadian respondents to round 1 of the PIAAC.

Understanding the challenges in the use of the GLMM framework when including a large number of process data, a more parsimonious approach to model response accuracy, response times and other process data from ILSAs is proposed by Klotzke & Fox (2019). By using Bayesian covariance structure modeling, the authors modeled the complex dependence structure of ILSA data and allowed the correction of between-subject differences in the dependence structure by including test-taker background variables (e.g, gender, computer experience, native speaker and education level).

### **Group of items/test-level analysis and group-level characteristics (layers 2, 3, 6)**

The majority of studies in this category (4 out of 7) analyzed test-taking engagement using process data. Besides the inclusion of test-taker characteristics for the analysis of process data from ILSAs, these papers also compare their results at country level.

Using PIAAC data, Goldhammer et al., (2016) derived indicators of test-taking engagement through response time thresholds at domain level (i.e., literacy, numeracy or problem solving), and explored subgroup differences (e.g., country or gender analysis). Pokropek (2016) also used the timing information from PIAAC data for detecting guessing behavior in a grade of membership modeling framework, providing a more precise estimation of group differences.

Test-taking engagement was also investigated using PISA 2009 data. Naumann (2015) used a GLMM approach to predict task performance by indicators of online reading engagement extracted from log-file data (i.e., navigation actions). In 2019, the same author investigated whether test-taker characteristics (i.e., comprehension skill, enjoyment of reading, and knowledge of reading strategies) would predict how much time students would devote to digital reading tasks. Naumann & Goldhammer (2017), in turn, evaluated the effects of response times in digital reading moderated by a person's skills and task demands in a GLMM framework. In these studies, the authors used a meta-analytical approach to compare the results from different countries.

By exploring process data from a high-stakes international language assessment, Lee & Haberman (2016) use a correlational analysis and summary statistics to study how students progressed in the test, their pace and management of time. Results were also compared across different test administrations and a selection of countries.

He et al., (2019), on the other hand, used a sequence-mining technique (i.e., the longest common subsequence) for an exploration of problem-solving strategies using process data from a selection of countries. Analyses of the differences in extracted strategies across different socio-demographic groups (i.e, gender, age, income, and familiarity with ICT) were also in the scope of this study.

### Discussion

In this chapter, we provided some examples of process data found in ILSAs, defined the kinds of information we get from such data, and proposed an ecological framework for their analysis. With a literature review of the first scientific publications in this field, we believe to have caught a glimpse of the seminal work as well as the current state-of-the art on the topic. We have also provided the reader with an overview of the common approaches and strategies these studies used in their empirical analysis of process data using ILSA data.

In this final section, we will discuss some considerations on the potential and limitations of process data in the context of ILSAs, how this study could add to the body of research in the field, as well as try to peek into the upcoming developments in process data analysis.

#### **The potential and limitations of process data in international large-scale assessments**

Given the great potential to offer insights on respondents' cognitive processes or attitudes in a technology-based assessment, empirical research on process data from ILSAs have been gradually emerging in scientific literature in the last decade. Log-file data from ILSAs, for example, can provide fine-grained recordings of the interactions between test taker and test items that was previously only available in small-scale experiments (e.g., cognitive interviewing, think-aloud protocols).

Log files also allow a deeper analysis of interactive items (e.g., recording of respondents clicking of buttons or links, selection of items in dropdown menus, dragging and dropping of on-screen objects with pointer devices like mice, copying and pasting texts) from computer-based assessments that was not possible on paper-based tests. Moreover, it can also be included in the scoring process to enhance the validity of the measured scores (Ramalingam & Adams, 2018; Engelhardt & Goldhammer, 2019).

Even though log files have driven significant contributions in the field of educational measurement with the improvement of psychometric models (e.g., joint modeling of response accuracy and response times), and exploratory analysis (e.g., data mining techniques), it is worth highlighting the importance of deriving valid process indicators from such files. Inferences of the latent process captured by these measures need to be justifiable both theoretically and empirically (Goldhammer & Zehner, 2017), which can be difficult when such indicators are constructed *ad hoc* from the available, sometimes severely limited log data (Kroehne & Goldhammer, 2018).

The complete picture of a test administration is not fully recorded in the key strokes and response times from log files. For instance, off-screen activities such as students' notes on paper and the use of a physical calculator are not captured in such technology. In this sense, Maddox et al. (2018) and Maddox (2017) argue for the use of complementary process data, such as eye-tracking or video-ethnographic observations, in ILSAs. However, there is a need to advance in the technology to expand the use of such process data in large-scale applications (e.g., speech or facial recognition systems to automatically derive process data from these tools). The argument for the validity of

## PROCESS DATA ANALYSIS IN ILSAs

such measures also needs be investigated. As a concern in the analysis of any response process, variations in the scores of such indicators can be attributed either to differences in the particular trait under investigation, or differential item functioning (Maddox et al., 2018).

### **How could the ecological framework contribute to the advancement of the field?**

We believe that the ecological framework for the analysis of process data can offer a visual mapping of the analysis strategies for a deep and comprehensive exploration of process data from ILSAs. The intrinsic multilevel structure of ILSAs datasets is the central feature of this conceptual model and will guide enthusiastic analysts towards a good start when working on process data. For instance, data pre-processing can be a better organized task when one set their research objectives based on the framework's layers. Extreme layers are linked with coarse grain levels of process data and more aggregation might be done.

The contribution of the ecological model to the next steps of the analysis plan after the data pre-processing stage may seem less clear, but this study adds to the body of knowledge on the process data analysis by showcasing the analytical approaches of over three dozen empirical studies that used process data from ILSAs. We categorized the works based on the ecological model with a view of what analyses' strategies the authors have proposed for each unit of analysis.

Even if the statistical approaches applied on these studies are not innovative and exclusive to a particular layer or even to process data analysis per se, they may inspire researchers to investigate a specific hypothesis/ analytical strategy based on the existing literature and the level of data that they have in hand. They may also instigate one to produce research publications that fill the gaps in the literature for the advancement of the field in mind. For instance, none of the studies from this literature review covered layer 4 (classroom and school context), even though there is data such as the information on opportunity to learn (OtL) collected by PISA that would allow researchers to explore and discover insights on this layer. As mentioned by De Boeck & Scalise (2019), data in the context of OtL reports classroom activities and practices, and exploring such relationships with process data might allow more reflection on a students' response process during the assessment.

Since the intention of this study was to present a sample of studies in the process data analysis, a systematic approach for the literature review is still in need. For example, we did not include in our review research reports such as OECD (2019) or Azzolini, Bazoli, Lievore, Schizzerotto, & Vergolini (2019), neither have we made use of traditional research datasets (e.g., ERIC, Web of Science) besides Scopus.

### **What is there to come?**

We believe that there is still a necessity for the development of tools to facilitate the data analysis. Perhaps one of the biggest hurdles for analysts enthusiastic about using process data in their studies is navigating through the torrent of data recorded in log files. Due to the lack of substantial theoretical foundation and/or software limitations, modern testing software is still unable to

## PROCESS DATA ANALYSIS IN ILSAs

identify which user actions are relevant when an examinee is interacting with an item, so they record as much as they can. The signal-to-noise ratio is just too low, at this point, and the recorded data is often cryptic (log files were originally designed for system maintainers to debug software, not for scientists looking for insight into the underlying processes of the human mind). In this context, tools such as the OECD LogDataAnalyzer (OECD, 2013) and the R package LOGAN (Reis Costa & Leoncio, 2019) are welcome contributions to the scientific community and have the potential of reducing the burden of the complex task of data management of such files.

One must also note that there is a growing concern regarding the ethical and privacy issues for the availability of such data for ampler use in research. Guidelines must be prepared to ensure the correct manipulation and use of such files. Thinking about regulatory issues and dilemmas for each unit of the process data analysis (e.g., de-identification of data at item or schools levels), the proposed ecological model may also be a useful tool in this task.

### References

- Azzolini, D., Bazoli, N., Lievore, I., Schizzerotto, A., & Vergolini, L. (2019). Beyond achievement. A comparative look into 15-year-olds' school engagement, effort and perseverance in the European Union. In E. Union (Ed.), *Entrepreneurship as Extreme Experience*. <https://doi.org/10.2766/98129>
- Chen, M. Y., & Zumbo, B. D. (2017). Ecological Framework of Item Responding as Validity Evidence: An Application of Multilevel DIF Modeling Using PISA Data. In *Understanding and investigating response processes in validation research* (pp. 53–68).
- Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). Statistical Analysis of Complex Problem-Solving Process Data: An Event History Analysis Approach. *Frontiers in Psychology, 10*, 486. <https://doi.org/10.3389/FPSYG.2019.00486>
- De Boeck, P., & Scalise, K. (2019). Collaborative problem solving: Processing actions, time, and performance. *Frontiers in Psychology, 10*, 1280. <https://doi.org/10.3389/fpsyg.2019.01280>
- Engelhardt, L., & Goldhammer, F. (2019). Validating Test Score Interpretations Using Time Information. *Frontiers in Psychology, 10*, 1131. <https://doi.org/10.3389/fpsyg.2019.01131>
- Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (No. 133; OECD Education Working Papers). <https://doi.org/10.1787/5jlzfl6fhxs2-en>
- Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics. *Large-Scale Assessments in Education, 5*, 18. <https://doi.org/10.1186/s40536-017-0051-9>
- Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626. <https://doi.org/10.1037/a0034716>
- Goldhammer, F., & Zehner, F. (2017). What to Make Of and How to Interpret Process Data. *Measurement: Interdisciplinary Research and Perspectives, 15*(3–4), 128–132.

## PROCESS DATA ANALYSIS IN ILSAs

<https://doi.org/10.1080/15366367.2017.1411651>

- Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers and Education, 91*, 92–105. <https://doi.org/10.1016/j.compedu.2015.10.018>
- Hahnel, C., Goldhammer, F., Kröhne, U., & Naumann, J. (2017). Reading digital text involves working memory updating based on task characteristics and reader behavior. *Learning and Individual Differences, 59*(October), 149–157. <https://doi.org/10.1016/j.lindif.2017.09.001>
- Hahnel, C., Goldhammer, F., Kröhne, U., & Naumann, J. (2018). The role of reading skills in the evaluation of online information gathered from search engine environments. *Computers in Human Behavior, 78*, 223–234. <https://doi.org/10.1016/j.chb.2017.10.004>
- Hahnel, C., Goldhammer, F., Naumann, J., & Kröhne, U. (2016). Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital text. *Computers in Human Behavior, 55*, 486–500. <https://doi.org/10.1016/j.chb.2015.09.042>
- Han, Z., He, Q., & von Davier, M. (2019). Predictive Feature Generation and Selection from Process Data in PISA Simulation-Based Environment: An Implementation of Tree-based Ensemble Methods. *Frontiers in Psychology, 10*, 2461. <https://doi.org/10.3389/fpsyg.2019.02461>
- He, Q., Borgonovi, F., & Paccagnella, M. (2019). *Using Process Data to Understand Adults' Problem-Solving Behaviours in PIAAC: Identifying Generalised Patterns across Multiple Tasks with Sequence Mining* (No. 205; OECD Education Working Papers).
- He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with N-grams. In *Quantitative Psychology Research: The 79th Annual Meeting of the Psychometric Society* (pp. 173–190). Springer International Publishing.
- Herde, C. N., Wüstenberg, S., & Greiff, S. (2016). Assessment of Complex Problem Solving: What We Know and What We Don't Know. *Applied Measurement in Education, 29*(4), 265–277. <https://doi.org/10.1080/08957347.2016.1209208>
- Hu, Y., Wu, B., & Gu, X. (2017). An Eye Tracking Study of High-and Low-Performing Students in Solving Interactive and Analytical Problems. *Journal of Educational Technology & Society, 20*(4), 300–311. <https://doi.org/10.2307/26229225>
- Huble, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In *Understanding and investigating response processes in validation research* (pp. 1–12).
- Klotzke, K., & Fox, J. P. (2019). Bayesian covariance structure modelling of responses and process data. *Frontiers in Psychology, 10*, 1675. <https://doi.org/10.3389/fpsyg.2019.01675>
- Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika, 45*, 527–563. <https://doi.org/10.1007/s41237-018-0063-y>
- Kroehne, U., Hahnel, C., & Goldhammer, F. (2019). Invariance of the Response Processes Between Gender and Modes in an Assessment of Reading. *Frontiers in Applied Mathematics and Statistics, 5*, 2. <https://doi.org/10.3389/fams.2019.00002>
- Krstić, K., Šoškić, A., Ković, V., & Holmqvist, K. (2018). All good readers are the same, but every low-

## PROCESS DATA ANALYSIS IN ILSAs

- skilled reader is different: an eye-tracking study using PISA data. *European Journal of Psychology of Education*, 33, 521–541. <https://doi.org/10.1007/s10212-018-0382-0>
- Lee, Y.-H., & Haberman, S. J. (2016). Investigating Test-Taking Behaviors Using Timing and Process Data. *International Journal of Testing*, 16(3), 240–267. <https://doi.org/10.1080/15305058.2015.1085385>
- Liao, D., He, Q., & Jiao, H. (2019). Mapping background variables with sequential patterns in problem-solving environments: An investigation of United States adults' employment status in PIAAC. *Frontiers in Psychology*, 10, 646. <https://doi.org/10.3389/fpsyg.2019.00646>
- Liu, H., Liu, Y., & Li, M. (2018). Analysis of Process Data of PISA 2012 Computer-Based Problem Solving: Application of the Modified Multilevel Mixture IRT Model. *Frontiers in Psychology*, 9, 1372. <https://doi.org/10.3389/fpsyg.2018.01372>
- Maddox, B. (2017). Talk and Gesture as Process Data. *Measurement: Interdisciplinary Research and Perspectives*, 15:3-4, 113–127. <https://doi.org/10.1080/15366367.2017.1392821>
- Maddox, B., Bayliss, A. P., Fleming, P., Engelhardt, P. E., Edwards, S. G., & Borgonovi, F. (2018). Observing response processes with eye tracking in international large-scale assessments: evidence from the OECD PIAAC assessment. *European Journal of Psychology of Education*, 33(3), 543–558. <https://doi.org/10.1007/s10212-018-0380-2>
- Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, 59(4), 439–483. <https://doi.org/10.1007/BF02294388>
- Mullis, I. V. S., & Martin, M. O. (2017). *TIMSS 2019 Assessment Frameworks*. Boston College, TIMSS & PIRLS International Study Center. <http://timssandpirls.bc.edu/timss2019/frameworks/>
- Naumann, J. (2015). A model of online reading engagement: Linking engagement, navigation, and performance in digital reading. *Computers in Human Behavior*, 53, 263–277. <https://doi.org/10.1016/j.chb.2015.06.051>
- Naumann, J. (2019). The skilled, the knowledgeable, and the motivated: Investigating the strategic allocation of time on task in a computer-based assessment. *Frontiers in Psychology*, 10, 1429. <https://doi.org/10.3389/fpsyg.2019.01429>
- Naumann, J., & Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learning and Individual Differences*, 53, 1–16. <https://doi.org/10.1016/j.lindif.2016.10.002>
- OECD. (2010). *PISA Computer-Based Assessment of Student Skills in Science*. OECD Publishing. <http://www.sourceoecd.org/education/9789264082021>
- OECD. (2013). *LogDataAnalyzer*. PIAAC Log File Website. <https://www.oecd.org/skills/piaac/log-file/>
- OECD. (2017). *PISA 2015 Technical Report*. OECD Publishing. <https://doi.org/10.1787/9789264255425-en>
- OECD. (2019). *Beyond Proficiency: Using Log Files to Understand Respondent Behaviour in the Survey of Adult Skills*. <https://doi.org/10.1787/0b1414ed-en>
- Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, 5(4). <https://doi.org/10.1186/s13643-016-0384-4>

## PROCESS DATA ANALYSIS IN ILSAs

- Pejic, A., & Molcer, P. S. (2016). Exploring data mining possibilities on computer based problem solving data. *SISY 2016 - IEEE 14th International Symposium on Intelligent Systems and Informatics, Proceedings*, 171–176. <https://doi.org/10.1109/SISY.2016.7601491>
- Pokropek, A. (2016). Grade of Membership Response Time Model for Detecting Guessing Behaviors. *Journal of Educational and Behavioral Statistics*, 41(3), 300–325. <https://doi.org/10.3102/1076998616636618>
- Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in Psychology*, 9, 2231. <https://doi.org/10.3389/fpsyg.2018.02231>
- Ramalingam, D., & Adams, R. J. (2018). How can the use of data from computer-delivered assessments improve the measurement of twenty-first century skills? In *Assessment and Teaching of 21st Century skills* (pp. 225–238). <https://doi.org/10.1007/978-3-319-65368-6>
- Reis Costa, D., & Leoncio, W. (2019). LOGAN: An R package for log file analysis in international large-scale assessments. *R Package*.
- Ren, Y., Luo, F., Ren, P., Bai, D., & Li, X. (2019). Exploring Multiple Goals Balancing in Complex Problem Solving Based on Log Data. *Front. Psychol*, 10, 1975. <https://doi.org/10.3389/fpsyg.2019.01975>
- Stone, J., & Zumbo, B. D. (2016). Validity as a Pragmatist Project : A Global Concern with Local Application. In V. Aryadoust & J. Fox (Eds.), *Trends in Language Assessment Research and Practice* (pp. 555–573). Cambridge Scholars.
- Teig, N. (2019). *Scientific inquiry in TIMSS and PISA 2015: Inquiry as an instructional approach and the assessment of inquiry as an instructional outcome in science* [University of Oslo]. <http://urn.nb.no/URN:NBN:no-74775>
- Ulitzsch, E., von Davier, M., & Pohl, S. (2020a). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, 73(S1), 83–112. <https://doi.org/10.1111/bmsp.12188>
- Ulitzsch, E., von Davier, M., & Pohl, S. (2020b). Using Response Times for Joint Modeling of Response and Omission Behavior. *Multivariate Behavioral Research*, 55(3), 425–453. <https://doi.org/10.1080/00273171.2019.1643699>
- Valdman, J. (2001). Log file analysis. In *Department of Computer Science and Engineering (FAV UWB), Tech. Rep. DCSE/TR-2001-04*. <http://www.kiv.zcu.cz/vyzkum/publikace/technicke-zpravy/2001/tr-2001-04.pdf>
- van Damme, D. (2014). How Closely is the Distribution of Skills Related to Countries' Overall Level of Social Inequality and Economic Prosperity? *OECD Education Working Papers*, 105(105), 1–23.
- Vista, A., Care, E., & Awwal, N. (2017). Visualising and examining sequential actions as behavioural paths that can be interpreted as markers of complex behaviours. *Computers in Human Behavior*, 76, 656–671. <https://doi.org/10.1016/j.chb.2017.01.027>
- Vörös, Z., & Rouet, J. F. (2016). Laypersons' digital problem solving: Relationships between strategy and performance in a large-scale international survey. *Computers in Human Behavior*, 64, 108–116. <https://doi.org/10.1016/j.chb.2016.06.018>
- Xu, H., Fang, G., Chen, Y., Liu, J., & Ying, Z. (2018). Latent Class Analysis of Recurrent Events in Problem-



## PROCESS DATA ANALYSIS IN ILSAs

Solving Items. *Applied Psychological Measurement*, 42(6), 476–498.  
<https://doi.org/10.1177/0146621617748325>

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, 71(2), 262–286.  
<https://doi.org/10.1111/bmsp.12114>

## PROCESS DATA ANALYSIS IN ILSAs

Table 1: A review of scientific papers, book chapters and working papers with empirical analysis of process data from ILSA.

Title	Authors-Year	Cited <sup>(a)</sup>	Operational variable	Domain	Assessment <sup>(b)</sup>	Layer <sup>(c)</sup>
Exploring data mining possibilities on computer based problem solving data	Pejic & Molce (2016)	1	Actions	Problem solving	PISA 2012	1
Latent Class Analysis of Recurrent Events in Problem-Solving Items	Xu et al. (2018)	3	Time and actions	Problem solving	PISA 2012	1
Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified Multilevel Mixture IRT model	Liu et al. (2018)	4	Time and actions	Problem solving	PISA 2012	1
Data mining techniques in analyzing process data: A didactic	Qiao & Jiao (2018)	4	Time and actions	Problem solving	PISA 2012	1
Statistical analysis of complex problem-solving process data: An event history analysis approach	Chen et al. (2019)	4	Time and actions	Problem solving	PISA 2012	1
Visualising and examining sequential actions as behavioural paths that can be interpreted as markers of complex behaviours	Vista et al. (2017)	11	Actions	Collaborative Problem Solving	ATC21S project	1
Collaborative problem solving: Processing actions, time, and performance	De Boeck & Scalise (2019)	N/A	Time and actions	Problem solving	PISA 2015	1
Exploring multiple goals balancing in complex problem solving based on log data	Ren et al. (2019)	N/A	Actions	Problem solving	PISA 2012	1
Predictive Feature Generation and Selection Using Process Data From PISA Interactive Problem-Solving Items: An Application of Random Forests	Han et al. (2019)	N/A	Time and actions	Problem solving	PISA 2012	1
Mapping background variables with sequential patterns in problem-solving environments: An investigation of United States adults' employment status in PIAAC	Liao et al. (2019)	2	Time and actions	Problem solving	PIAAC	1, 5
Identifying feature sequences from process data in problem-solving items with N-grams	He & von Davier (2015)	18	Actions	Problem solving	PIAAC	1, 6
Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving	Greiff et al. (2015)	62	Actions	Problem solving	PISA 2012	1, 6
Laypersons' digital problem solving: Relationships between strategy and performance in a large-scale international survey	Vörös & Rouet (2016)	2	Time and actions	Problem solving	PIAAC	2
Observing response processes with eye tracking in international large-scale assessments: evidence from the OECD PIAAC assessment	Maddox et al. (2018)	3	Time and actions	Literacy, numeracy, and problem solving	PIAAC	2
How can the use of data from computer-delivered assessments improve the measurement of twenty-first century skills?	Ramalingam & Adams (2018)	6	Actions	Reading	PISA 2012	2

## PROCESS DATA ANALYSIS IN ILSAs

Title	Authors-Year	Cited <sup>(a)</sup>	Operational variable	Domain	Assessment <sup>(b)</sup>	Layer <sup>(c)</sup>
How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items	Kroehne & Goldhammer (2018)	8	Time and actions	Context questionnaire	PISA 2015	2
Cognitive diagnosis modelling incorporating item response times	Zhan et al. (2018)	23	Time	Mathematics	PISA 2012	2
Validating Test Score Interpretations Using Time Information	Engelhardt & Goldhammer (2019)	N/A	Time	Literacy and reasoning	PIAAC	2
The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment	Goldhammer et al. (2014)	145	Time	Reading and problem solving	PIAAC	2, 3
Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics	Goldhammer et al. (2017)	7	Time	Literacy, numeracy, and problem solving	PIAAC	2, 3
Invariance of the Response Processes Between Gender and Modes in an Assessment of Reading	Kroehne et al. (2019)	1	Time	Reading	PISA 2009	2, 3
Reading digital text involves working memory updating based on task characteristics and reader behavior	Hahnel et al. (2017)	3	Actions	Reading	PISA 2012	2, 3
Using Response Times for Joint Modeling of Response and Omission Behavior	Ulitzsch et al. (2020a)	3	Time	Numeracy	PIAAC	2, 3
A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response	Ulitzsch et al. (2020b)	N/A	Time	Mathematics	PISA 2015	2, 3
All good readers are the same, but every low-skilled reader is different: an eye-tracking study using PISA data	Krstić et al. (2018)	5	Time and actions	Reading	PISA	2, 3
An eye tracking study of high- and low-performing students in solving interactive and analytical problems	Hu et al. (2017)	9	Time and actions	Problem solving	PISA	2, 3
Talk and Gesture as Process Data	Maddox (2017)	8	Time and actions	Literacy, numeracy, and problem solving	PIAAC	2, 3
The role of reading skills in the evaluation of online information gathered from search engine environments	Hahnel et al. (2018)	20	Actions	Reading	PISA 2009	2, 3
Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital text	Hahnel et al. (2016)	56	Actions	Reading	PISA 2009	2, 3
Bayesian covariance structure modelling of responses and process data	Klotzke & Fox (2019)	N/A	Time and actions	Numeracy and literacy	PIAAC	2, 3
Using Process Data to Understand Adults' Problem-Solving Behaviours in PIAAC: Identifying Generalised Patterns across Multiple Tasks with Sequence Mining	He et al. (2019)	1	Actions	Problem solving	PIAAC	2, 3, 6

## PROCESS DATA ANALYSIS IN ILSAs

Title	Authors-Year	Cited <sup>(a)</sup>	Operational variable	Domain	Assessment <sup>(b)</sup>	Layer <sup>(c)</sup>
Grade of Membership Response Time Model for Detecting Guessing Behaviors	Pokropek (2016)	3	Time	Numeracy	PIAAC	2, 3, 6
Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands	Naumann & Goldhammer (2017)	13	Time and actions	Reading	PISA 2009	2, 3, 6
Investigating Test-Taking Behaviors Using Timing and Process Data	Lee & Haberman (2016)	13	Time and actions	Reading	International Language Assessment	2, 3, 6
Test-taking engagement in PIAAC	Goldhammer et al. (2016)	23	Time	Literacy, numeracy, and problem solving	PIAAC	2, 3, 6
A model of online reading engagement: Linking engagement, navigation, and performance in digital reading	Naumann (2015)	39	Actions	Reading	PISA 2009	2, 3, 6
The skilled, the knowledgeable, and the motivated: Investigating the strategic allocation of time on task in a computer-based assessment	Naumann (2019)	N/A	Time and actions	Reading	PISA 2009	2, 3, 6

Notes: (a) Number of citations that the work received between November 2019 and January 2020 in the Google Scholar database. (b) We identify the cycle and the type of assessment to which the items belongs, not necessarily from where the process data were extracted. That is, we included studies in this review with process data from the assessment's main study, field trial or new administration with a different target population. (c) Indication of the correspondent layer of the ecological framework for the analysis of process data in ILSA.