How Students Behave While Solving Critical Thinking Tasks in an Unconstrained Online Environment: Insights from Process Mining

Anastasia Beliaeva

Received in October 2023

Anastasia Beliaeva — Intern Researcher at the Center of Psychometrics and Measurements in Education, HSE University. Address: 16/10 Potapovsky lane, 101000 Moscow, Russian Federation. E-mail: belyaeva.a.u@yandex.ru. ORCID: https://orcid.org/0000-0002-4855-4390

Annotation

To learn successfully with the use of various internet resources, students must acquire critical thinking skills that will enable them to critically search, evaluate, select and verify information online. Defined as Critical Online Reasoning, this complex latent construct manifests itself in an unconstrained online environment and is measured on two levels: students' work product (an essay) and the process of task completion (online behaviour patterns). This research employs process mining techniques to investigate the possibility of distinguishing between students' successful and unsuccessful attempts to take the test. The findings of the work were gained on generalised behaviour patterns from the process mining algorithm deployed on two groups of students (63 low performing and 45 high performing students). Divided by the work product score, the two groups exposed some differences in their online behaviour, with the high performers showing more strategic behaviour and effective search and use of information online. However, the research has also shown the downside of process mining as a tool for generalisation of process patterns.

Keywords

critical online reasoning, process mining, event logs, process patterns

For citing

Beliaeva A.Yu. (2024) How Students Behave While Solving Critical Thinking Tasks in an Unconstrained Online Environment: Insights from Process Mining. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3 (1), pp. 4–26. https://doi.org/10.17323/vo-2024-18051

Critical thinking (CT) and its manifestation in education has been in the scope of scientific research since the beginning of the previous century [Dewey, 1910]. With the advent of the new era of information and computer technology, enabling people to access any type of data online, there was a surge of works on critical online reasoning (as it was introduced by Zlatkin-Troitschanskaia et al. [2020]) or civic online reasoning (as it was coined by McGrew and Wine-

burg) or as the acronym COR, which is a branch of critical thinking related to working with data online, be that searching, browsing, checking the reliability of the source, or a similar activity [McGrew et al., 2018]. In the current study, critical thinking in an online environment is the ability of university students to analyse statements, assumptions and arguments, build causal relationships, select logically correct and convincing arguments, find explanations, draw conclusions, and form their own position in solving problems in an online environment.

It is considered that there is a particular need to develop this skill among university students, as those of them who have not acquired the strategies for successful search may struggle with various tasks they are required to fulfil, hence their low performance at university [Liu et al., 2016].

To form this skill, it is necessary to explain the nature of the successful strategies to students, and then check whether they use those strategies effectively. Hardly is it possible to develop a firm habit of searching without providing relevant feedback. Prompt feedback could be of particular value as being provided shortly after testing, it can engage and motivate underperforming students to catch up on the subject and with the rest of the group [Lightbown, Spada, 2021].

To this end, testing is to be implemented and students' attempts are to be analysed. To do so, their digital footprints (log-files) are to be recorded into a log journal for further in-depth analysis. To draw valid conclusions about students' COR, test designers try to save rich log files [Padilla, Benítez, 2014]. Yet, with a growing number of logs to be analysed it is becoming a daunting task for the assessors, especially if testing is carried out on a large scale.

However, generalising a plethora of answers into less intricate yet common patterns may substantially facilitate the interpretation of results. Instead of giving final marks, assessors could analyse the patterns to differentiate examinees who genuinely tried to solve the task but lacked the necessary skills from those who were unmotivated and did not put in their best effort [Ulitzsch et al., 2022]. Reasonably, those who lacked the skill but not the motivation could be advised on a better approach.

Looking at the issue from a different perspective, universal competencies, like critical thinking, are composite latent constructs [Mislevy, 2018)]. To evaluate such constructs, when developing an assessment model, it is necessary to describe the situation in which we are going to place the test taker in order for him or her to demonstrate the skills reflected in the definition of the respective construct. It describes the very environment in which we will evaluate (e.g., an essay, simulation, or a game) and the actions that will

be performed in this environment — the activity space. As a rule, this model includes the following [Mislevy, 2013]:

1. detailed "targets" for assessment, the observed variables detailed to the required extent (what we want to observe);
2. the type of stimulus or materials that will be used to observe them;
3. a description of what exactly the person being evaluated will be asked to do during the evaluation;
4. a description of the elements that must be present in the task so that the evaluator can demonstrate the actions that we want to observe;
5. the elements that will affect the complexity of tasks.

Thus, assessment of complex latent constructs is based on collected behavioural characteristics gained in the process of performing actions (process data) in the assessment environment. Fine tuning of a test is by far more profound if done by looking into the rich information that is embedded in the actions of the test takers. What is more, the analysis of patterns can also lend weight to theories about construct manifestation.

Literature review

Automated pattern extraction is widely used in the sphere of closed-environment tasks. In particular, much research on behavioural patterns is conducted using datasets from OECD problem-solving tasks in various domains of knowledge [Ulitzsch et al., 2022]. There are successful attempts to generalise patterns and to examine incorrect behaviour as a specific area of interest. For example, a study concluded that incorrect answers can be caused by a lack of strategy or failure to implement it successfully [Stadler et al., 2019]; according to Ulitzsch et al. [2022] literature review, most scholars agree that wrong responses involve behaviour that is more imbalanced and tangled, with more deviations and fewer similarities than that of the correct ones. Also, longer time spent on a task positively correlates with the grade obtained [Eichmann et al., 2020; Stadler et al., 2019; Tang et al., 2020; Ulitzsch et al., 2022] since spending time implies making effort. However, according to Eichmann et al. (2020) the majority of complex problem-solving tasks analysis only scratch the surface of sequence analysis per se, as research of full patterns is outnumbered by studies considering either frequency of the actions or time spent on task completion. To the best of our knowledge, little research is conducted on students acting in an *open unconstrained online environment* and even fewer attempts are made to analyse the patterns through process mining.

Although COR has been studied to the best of 25 years, thus far it is still onerous to automatically determine and process the patterns of students' behaviour in the process of information search [Schmidt et al., 2020]. Extensive research has been carried out on whether or not COR can be tested using multiple-choice or other classical testing techniques. The majority of authors admit that COR cannot be constricted to the classical multiple-choice test means [Griffin et al., 2012; Weber et al., 2019; Zlatkin-Troitschanskaia et al., 2020; Molerov et al., 2020; Zlatkin-Troitschanskaia et al., 2021]. However, in some cases, classical testing and open search assessment can be combined to cover the skill as thoroughly as possible [Tarasova, Orel, 2022].

To illustrate the existing research on evaluating COR through automation, I conducted a comprehensive literature review. To this end, an approach proposed by Kitchenham and Charters [Kitchenham, Charters, 2007] was adopted, which included a systematic search for primary papers with the help of key searching. The key words were either Critical Online Reasoning or Civic Online Reasoning. I used three scientific article aggregators, namely Google Scholar, ResearchGate and Connected Papers. The multiple search engines were employed in order to comply with the triangulation method of checking information.

The results of the search included fewer than 100 works for Critical and Civic Online Reasoning. The next stage was to exclude all the papers that were not concerning the practical assessment of the process of doing the test (searching the web and writing the final work product) (e.g. works devoted to theoretical description of the construct). The goal was to identify works describing indicators, logging of the process and aggregation of the data obtained during a COR test as well as its evaluation and insights gained from such aggregation. As most of the papers dealt with teaching the strategies of critical web-search, developing and validation of theoretical frameworks for COR assessment, solely product assessment or theoretical reasoning of successful and unsuccessful COR manifestation, the number of articles relevant to the current study shrank to five. The scope of this article does not include expert assessment or merely descriptive analysis of the data; thus, such papers were also screened out during the search stage.

The selected research papers were then scrutinised in order to register the differences between analysis approaches and the results obtained, all of which are shown in Table 1. However, from the table it is apparent that these studies (except the one by Schmidt et al. 2020) did not attempt to explore the patterns of behaviour, but rather described and summarised some of its prominent features (like one common action in a group leading to success or failure).

The aggregation of behaviour and analysis of the whole pattern of COR task completion is limited to only [Schmidt et al., 2020]. In their research, students' actions as well as eye-fixations were simultaneously recorded to analyse the difference between low-achievers and high-achievers. The study was carried out on 32 participants, all of whom were undergraduate students. The authors initially gathered much process information; however, some of it (mouse clicks, keyboard strokes) were discarded during the analysis stage due to its complexity. Data visualisation was executed with special software called PAFnow. In the end, the researchers were not fully satisfied with pattern graphs derived from the data, as students' behaviour, being quite varied, was hardly susceptible to generalisation. Crude division of students, namely either low or high performers, may underlie the ambiguity of resulting patterns and account for poor model fit.

However, at this stage of research it is crucial to take one step further. Apparently, the majority of research on COR behaviour analysis comes down to separate action descriptions. Not much is done in order to dig deeper and see the bigger picture, in other words, to try to derive patterns inherent to the whole cluster of test-takers.

Process analysis as a tool of *generalising* the behaviour (i.e. drawing on common sequences of steps particular to a group and identifying the staple of this particular group opposed to others) has been successfully implemented in various adjacent fields, ranging from students behaviour analysis in online courses in order to predict whether or not a students is likely to pass the exam [Arpasat et al., 2021] to business workflow analysis in order to identify bottlenecks of business processes with a view of speeding up trade [Benevento et al., 2022]. As opposed to description, *generalising* requires systematic evaluation and aggregation of all behaviour encountered in the order it happens, which is barely possible to pinpoint manually. Thus, the core division between description and generalisation is that the latter is a powerful method of data analysis rather than a superficial manual registration.

As it is clear from the literature review, not much knowledge has been gained so far about the patterns of behaviour at different levels of critical online reasoning. The sparse research conducted has yet been unable to satisfactorily visualise and describe commonalities in students' behavioural patterns. Yet there are multiple examples of research that successfully implement process mining analysis and lend valuable insights into the underlying behaviour of the subjects. Thus, this research is aiming at answering the following research question (RQ):

1. What distinguishable features of low and high performing students can be identified with regard to duration, number, type, and order of steps in the test-taking process?

Table 1. **Analysis of COR manifestation in literature**

Authors	Method	Analysis	Results
Wineburg & McGrew [2017]	Participants (10 PhD historians, 10 fact-checkers, and 25 bachelor students) were asked to verbalise their thoughts while progressing through the task of website evaluation	Assessment rubrics were developed and two raters were employed and tested on interrater reliability (Cohen's Kappa); COR scores of groups were analysed with Mann-Whitney criterion	The authors concluded that the major attribute to higher COR score is in "taking bearings" and "lateral reading", meaning the ability to plan further analysis and omitting all irrelevant information, respectively
McGrew et al. [2018]	Three groups of participants (405 mid-school, 308 high-school, and 141 college students) were handed out paper-pencil tests or sent to Google Forms to write a short answer as to why the website is reliable/unreliable etc. They were instructed to browse the web to draw their conclusions	Rubrics for assessment were designed, revised and checked on interrater reliability afterwards (Cohen's Kappa).The scores of the groups were compared	Common fallible behaviour was observed and described in the discussion section. For example, when instructed to check the reliability of a site, students did not implement a "fact checking" strategy, but rather stayed on the initial website, never leaving it in search of extra information
Weber et al. [2018]	Two waves of surveys were conducted with 3816 and 769 undergraduates from different domains. The authors wanted to establish a relationship between the information seeking strategies used (advanced/ traditional/ basic) and the academic outcome of the subjects	OLS regression was implemented with a dependent variable of academic outcomes and the predictor — reported information-seeking strategies	One insight was about a significant difference of dominant strategies across domains of studies (e.g., medicine students deploy advanced strategies less often); another fact was that while students' strategies were becoming more elaborate over time, the basic strategies did not change
Nagel et al. [2020]	Students were given three 10-minute tasks, in which they were expected to use a browser in order to answer the task question in the form of a short essay. A subsample of 45 students' works was taken to analyse the process and product. The used sources were evaluated by independent raters	Interrater reliability (Cohen's Kappa) was checked for coding resources used by the students (the log-data). A t-test was conducted to see if there was a significant difference between the students who visited extra websites and those who did not	The students who had visited additional websites (not specified in the task itself) gained a significantly higher COR score
Schmidt et al. [2020]	32 purposely selected participants underwent a COR test that implied assessing two websites on their credibility. 10 minutes were reserved for this. Their actions were recorded as well as their eye-movements tracked for further analysis	Process mining was implemented with special software. Only part of the log-files were used in the process mining analysis. Eye-fixations were also analysed with process mining analysis. Latent class analysis (LCA) was carried out on the low and high performing groups of students	The various patterns students left during task completion are difficult to present in a visually clear way. The plots are too intricate with details for both logs and eye-fixation patterns. LCA proved that through process behaviour analysis one can classify students as low or high performers

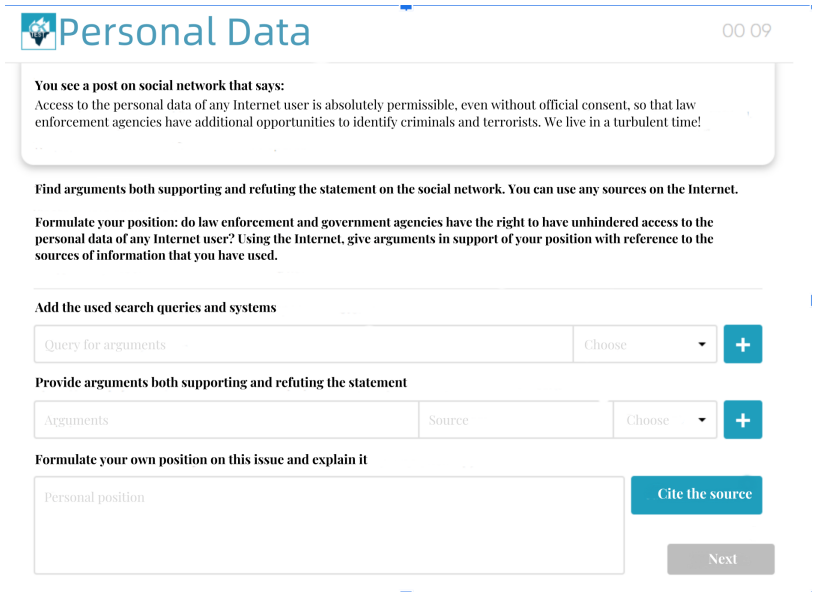
2. Research method

2.1. Sample and Procedure

The testing was a part of the course "Economic thinking". The data was collected in December 2021. The sample consisted of 330 students, who took the test under the following conditions: the students were given a link leading to a software that could trace their

online behaviour and register steps in a log journal. Figure 1 provides an overview of this part of the test (the test was conducted in Russian; the figure is a translated version of the main task screen of the test). There is a task and several fields for answers, which include the essay (the lowest field), the argument and the links (in the middle), the query and the browser (at the top). The duration of task completion was recorded as well as students' work in the aforementioned fields. The test was available for completion for one week; however, the students were instructed that it was preferable to complete the test within 1 hour and 30 minutes. Nevertheless, the system did not terminate their internet session if the time limit was exceeded (see Limitations). Another requirement of the task was to fill in arguments for and against before writing an essay. However, as it will be seen later, not all students came up with the requested number of arguments before commencing with the essay.

Figure 1. **Task virtual environment**



In total, students generated over 23.000 events while solving the unconstrained task. These numerous logs contained timestamps of each action. The software could register the following actions: typing in the essay field, the argument field, and the query one; leaving references in the essay and argument fields; choosing a web browser; adding or deleting an argument; adding or deleting a browser name; stating the type of argument (for/against). In total, the dataset encompassed 80 unique actions that the students made.

Each action was saved by a respective name. For example, if a student was typing a name of the source for their first argument, it appeared like "ZoneArgSource1" in the dataset, where "Zone" was a unique prefix for all actions, "Arg" stood for argument and "Source1" denoted the first source mentioned in this field. Table 2 provides an overview of a part of the dataset related to one test taker, namely the number of times a certain action was taken by him/her (*Action_num column*), the exact action performed (*Action_zone column*), the type of the action (pause, resume or click) (*Action_type column*), the time when the action started (in Python timestamp format) (*Action_time column*) and the duration of the action (in seconds) (*Duration column*), respectively. The ID of the student was removed from this table.

Table 2. **Example dataset**

Action_num	Action_zone	Action_type	Action_time	Duration
1	ZoneApp	Pause	1639843747	178
2	ZoneApp	Resume	1639843925	153
3	ZoneReqEngine1	Click	1639844078	204
4	ZoneReqEngine1	Click	1639844282	2
5	ZoneReqText1	Click	1639844284	2
6	ZoneApp	Pause	1639844286	30
7	ZoneApp	Resume	1639844316	13
8	ZoneArgSource1	Click	1639844329	4
9	ZoneArgRadio1	Click	1639844333	1
10	ZoneArgRadio1	Click	1639844334	8
11	ZoneArgSource1	Click	1639844342	1
12	ZoneArgText1	Click	1639844343	6
13	ZoneApp	Pause	1639844349	65
14	ZoneApp	Resume	1639844414	4
15	ZoneEssayText	Click	1639844418	21

2.2. Instrument

To evaluate critical thinking, a two-part test using Evidence-Centred Design (ECD) [Zieky, 2014] was developed. Both parts are based on the same theoretical framework and designed to cover all parts of critical thinking: analysing arguments, developing sound arguments and understanding causation and explanation (see for details [Tarasova, Orel, 2022]. Statistical analysis for the first part of the test was carried out in order to ensure that the test is valid. Cronback's alpha (0.59), fit statistics (RMSEA < 0.05; 0.85 < OutFit & InFit

< 1.15) and dimensionality analysis (eigenvalue of the first contrast = 1.97) demonstrated adequacy of the test.

In the second part, students are presented with a dilemma with no unambiguously correct answer (e.g. whether or not the government should have access to personal data in order to lower crime rates) and are instructed to state their opinion using an unconstrained online environment, where they can find relevant information for their argumentations and essays. The software the students are using can register their footprints, and they are also expected to fill in a form with required information about their work e.g. the resources used, the queries made, and the browsers surfed.

2.3. Data Analysis Approach

As opposed to the previous studies (e.g. [Schmidt et al., 2020]), I chose to split the data set not into three rather than two classes (high and low performers), but into three(high, average, and low). When only splitting into two groups, the contrast is vague; however, it is desirable to amplify it. It is rather burdensome to generalise students' behaviour as it is highly heterogeneous. On the other hand, splintering the database into three classes and afterwards keeping only high and low performers and dropping the average ones might substantially facilitate the generalisation process, thus providing more insights into class-specific strategies.

Thus, the score from the first part of the CT test was used to divide the students into three groups. To this end, I calculated the mean and standard deviations of the scores. The first standard deviation to the left and right from the mean score contained the average group. Starting from the second standard deviations and everything to the left of it was the low performing group, and to the right was the high performing one. For the low performing group, the score was from 0 to 12 points for the CT test, for the average one from 13 to 25, and for the high performing one from 26 to 38 (the maximum was 40 points, which no-one scored). As a result, there were 45 students with high scores, 63 with low scores, and 145 with average ones. The average score students (145 students) were excluded from the analysis altogether as their behaviour was out of the scope of the article research questions. In the next sections, only the low (63 students) and high (45 students) performing groups' patterns will be scrutinised.

Since the task was not formally controlled in terms of duration, some outliers were contaminating the dataset. For instance, there was a pattern that lasted for 7 days. Apparently, it is not possible to work for this length of time without interruption. Consequently, such works that exceed a reasonable threshold of time needed to take the test were deleted from the dataset.

To analyse each of the two groups separately and extract patterns, I used ProM[1] — an open source software for process discovery, and to look into the common patterns of each cluster in particular, I employed a Heuristics miner and a C-net model. Schmidt et al. [2020] performed their pattern visualisation in a similar software. Hence it is feasible to implement such programmes for this type of analysis.

The Heuristics miner and C-nets are expected to be suitable process discovery as they can deal with less structured behaviour [Rozinat, Aalst, 2009]. In essence, they work on the premise that the most frequent behaviours should be given precedence over least frequent ones to form a connection with the previous step. Given that the first step is unified within all the data set, next steps are calculated based on the so-called dependency graph, which indicates how certain the module is about the next step. An 80% threshold is used to suppress noisy behaviour. More on Heuristics miner and C-nets can be read in Weijters and Aalst [2006].

In the first phase, I carried out a general analysis of the data to determine the average amount of time spent on the task, the number of steps students took in general, the number of sources used, and other variables.

For further analysis of the data with ProM, I had to remove the noise, which implies deleting not popular behaviour, as such a large number of log files in a poorly structured behaviour will lead to poor performance of the miner [Aalst, 2016; Eichmann, 2020]. Due to the fact that the students were not limited by formal requirements for the task, the patterns are diverse. By reducing noisy unpopular behaviour, it is possible to increase the amount of information that can be potentially extracted from the given data [Aalst, 2016]. ProM offers a threshold to suppress slight deviations within the groups; the recommended level of filtering is 80% [Ibid.], which this work will adhere to.

The data was also processed and cleaned of the outliers. For instance, for seven students, the test procedure lasted more than three hours, and, though no formal restrictions of time were placed on the subjects of the test, these examples were regarded as noisy and thus removed from the dataset.

3. Research results

Regarding the RQ, the data was scrutinised using ProM in an attempt to establish common patterns for different levels of COR. The dataset was split into three clusters in accordance with the framework and the students' COR score. Then only those catego-

[1] ProM can be freely downloaded from https://promtools.org/

rised as high (45 instances) and low (63 instances) performers were kept to process mining.

To conduct an in-depth analysis of the general statistics and scrutinise the datasets before drawing graphs, a library for Python (version 3.9) for panel data analysis called pandas (version 2.0.3[2]) was used. To visualise the process of solving the task, ProM (version Lite 1.3) Heuristics miner was employed. The two clusters were analysed separately.

3.1. Low performers

To start with, the raw data was analysed using pandas to aggregate the dataset for descriptive statistics. On average, low performers only used 13 unique actions in the app, which resulted in most of them submitting only 1–2 arguments and rarely editing anything, be that a source, request, or text. Having analysed the dataset, it has also become apparent that the students spent on average 12 minutes working on their task. The mean time the students required to search online and read the websites' information amounted only to four minutes. Considering the overall time on submission, the students spent one third of their time surfing the web and most of the remaining time on actually submitting the answer in the answer form. In particular, it took students 12 switches between the application and the internet to reinforce their answer. What is also peculiar is the fact that only 14% of students started from leaving the app in search of an answer, with the vast majority of them commencing straight with filling in the form of the app. However, the most popular first step was to write a request — 44%.

To provide more insights into the behaviour in the graphical form, Heuristics miner was employed and graphs were extracted from the patterns. After fine tuning, 73 directly-follow steps (a pair of steps following one another directly) out of 259 instances fit the diagram process. The more instances fit, the more precise the model is; however, to enable the miner to draw sound and not spaghetti-like models, the subset of 63 students had to be degraded into a less diverse one. Thus, prior to the analysis, a Simple Heuristics filter preserved only 80% of all activities (in terms of how frequent each activity was); if some students' activities in the patterns were rare, they were removed from the dataset. As a result of the preparation, the dataset shrank to 43 subjects. Figure 2 shows a general pattern the miner generated. For this general pattern, only activities that happened at least seven times were included, while all the others were ignored.

In the graph, there are 10 boxes, each referring to a certain action in the pattern. For example, at the top there is "ZoneApp" which

[2] Can be downloaded freely from https://pandas.pydata.org/

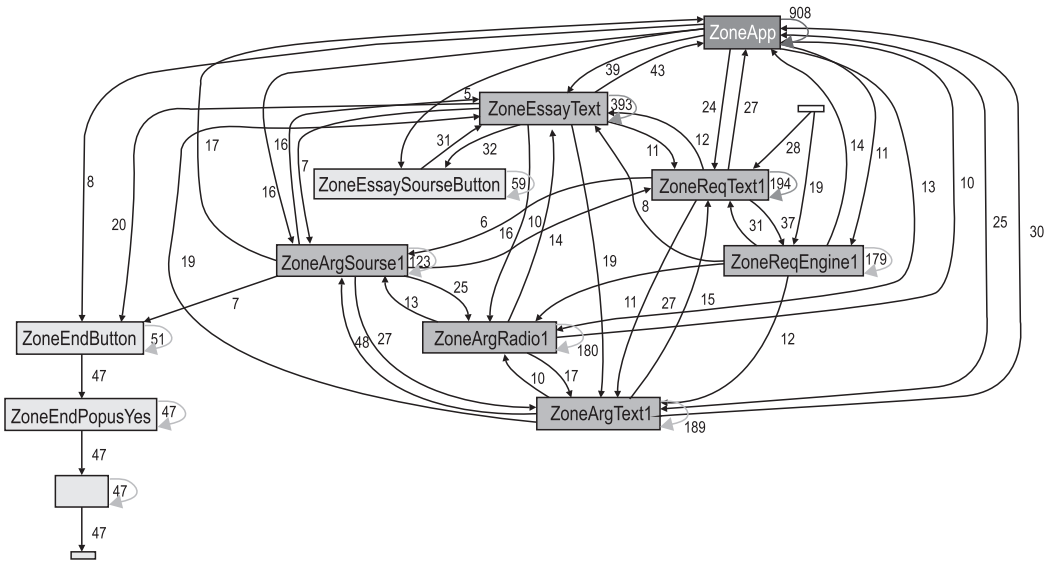
denotes the fact that the students left the app 908 times to search online. All the other boxes can be deciphered in the same fashion: "ZoneEssayText" means typing the essay, "ZoneEssaySourceButton" is adding a new source to the essay by clicking on a "+" button from Figure 1, etc. Noticeably, the number in the names of the boxes (e.g. "ZoneArgSource1") indicates that all students made this action, i.e. wrote their first argument source. It is possible, though, that there were instances of students submitting their second argument as well; however, since there is no "ZoneArgSource2" in the diagram, it implies that this behaviour was not popular in this cluster.

The intensity of the blue colour denotes the popularity of the actions; the figures accompanying each action box mean how many times this particular action was performed by all the participants. As it can be seen from the infographics, the majority began their work in the app not by checking the information necessary to accomplish the task online, but by filling in the form straight away. The participants did leave the app to check the information, though they did not do that prior to writing their answer in the form. Another helpful technique of studying such diagrams would be in observing the loops (when one activity directly follows itself) in order to see the intensity of students' work and compare it within different activities performed. For instance, students oscillated between the app and the internet 908 times, which can be seen as "ZoneApp" action (the semi-circular arrow means that the activity was terminated and then started again); at the same time, they started writing the essay and then stopped again 393 times in "ZoneEssayText" (with the same semi-circular arrow in the diagram). The rectangles with no captions denote the beginning and end of the whole process (with the latter following "ZoneEndPopupYes").

It may also be useful to consider the popularity of certain steps happening in combination. For example, there were 39 instances of students writing the essay ("ZoneEssayText") after coming back from the internet search (see the light blue arrows pointing from "ZoneApp" to "ZoneEssayText"). At the same time, they wrote their argument only 30 times after checking the information online (the arrow from "ZoneApp" to "ZoneArgText1"). This can provide a useful insight into the strategies the students used in order to solve the task (see the Discussion).

Yet, such a diagram might seem overwhelming because of the number of various steps and deviation between them. Another notation available in ProM is a Causal net (C-net) graph. Figure 3 offers an example. Here, the bindings (the light blue line connections) refer to a chain of activities that are paired or tripled together, meaning they happen concurrently. It provides a more intimate knowledge about the popularity of certain combinations of steps

Figure 2. **Low performers' general pattern**

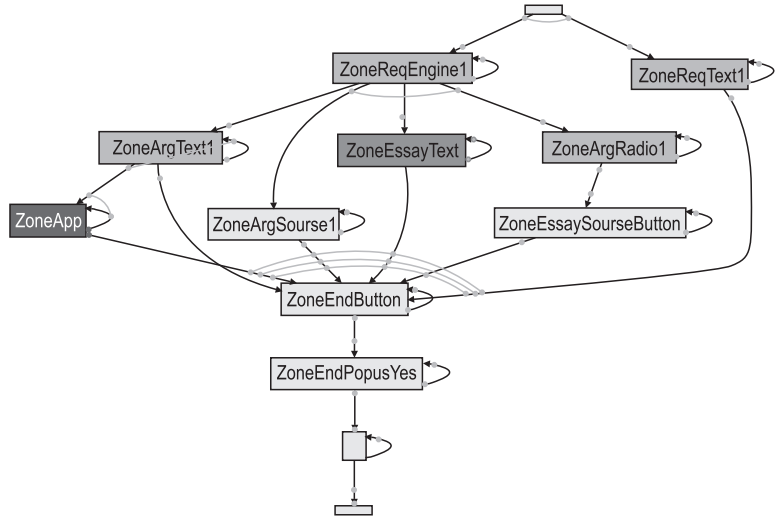


throughout the process of task completion, which deepens understanding of the pattern.

One thing to be cautious about is not to interpret the graph as showing the steps from the beginning to the end. Indeed, there is an empty rectangle at the top and two empty rectangles at the bottom, which indicate the beginning of the work and the end of it. However, the blue dots on every binding indicate that the process could either flow "down" or "up" the diagram. The bindings, on the other hand, show that the next step in the process could be either of the actions connected by the line. For example, after filling in the query ("ZoneReqEngine1"), students would commonly go to write the argument ("ZoneArgText1"), essay ("ZoneEssayText1"), argument source ("ZoneArgSource1"), or marked the argument as for/against ("ZoneArgRadio1") in no exact order.

A C-net graph does not provide any frequencies of steps (apart from the colour coded squares, where the darker blue indicates the more popular actions). Yet it is much more concise than the previous graph in Figure 2 and can provide some understanding of common steps completed in combination. Figure 1, in contrast, provides a "bird-eye view" of the whole process flow. However, it is apparent that the graph is too intricate, with many loops and spontaneous steps students take. Thus, Figure 3 can be considered more informative in terms of the process commonalities for all students in the group (namely, the cluster activities they chose to do together).

Figure 3. **C-net for low performers**



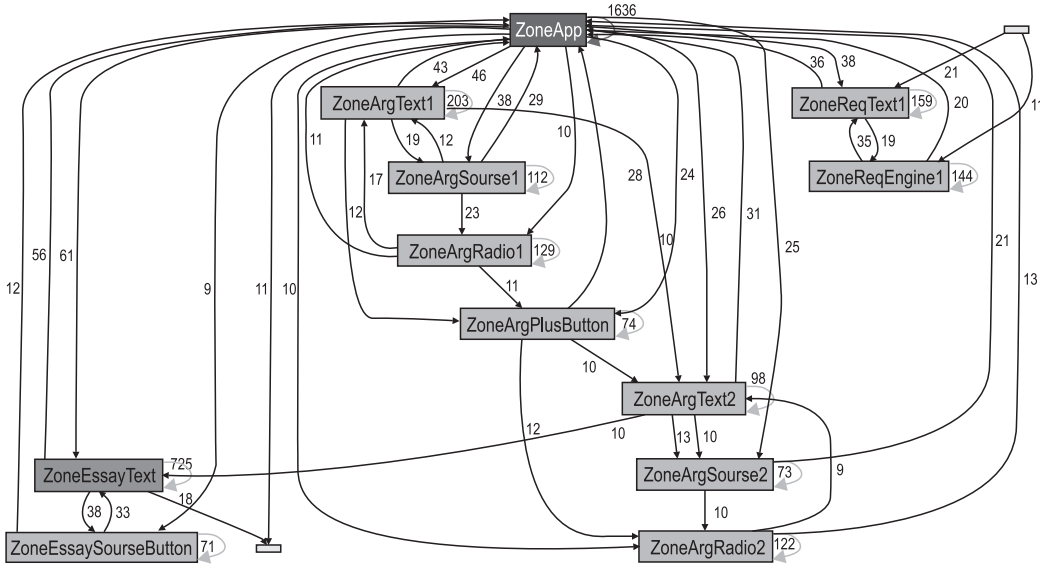
3.2. High performers

As it was done for the previous group, the high performers data was analysed using pandas to aggregate the dataset for the descriptive statistics. They generated 20 unique actions on average; this resulted in their submitting the requested number of arguments, which was at least two (see Sample and Procedure). The high-flyers switched between the app and internet 35 times on average. They submitted their tasks after 20 minutes of continuous work on average. The high achievers also spent more time in the app, either working on the essay, arguments or links, which was 13 minutes, roughly 0.65% of the whole work process. Interestingly enough, only 13% of students started from going out of the app in search of an answer, with the majority filling in the request form — 47%.

To analyse the graphical representation of patterns of the high performing cluster, Heuristics miner was used. Similarly to the previous analysis, this one was carried out on a filtered dataset with only 38 instances included (an 80% threshold served only the most common behaviour observed). Figure 4 illustrates the derived pattern. After fine tuning, 74 directly-follow steps (a pair of steps following one another directly) out of 372 instances fit the diagram process.

This graph shows the frequency of certain activities, colour-coded in shades of blue. As was the case with the low performing cluster, here the deeper the shade, the more occurrences were registered in the dataframe. Apparently, here the most popular activities were “ZoneApp” (leaving the app to search online) and “ZoneEssayText” (writing the essay). As with the low performers’ analysis, it does not provide any linear representation of steps taken by the students, making the sequencing opaque. In other words, one can

Figure 4. **High performers' general pattern**



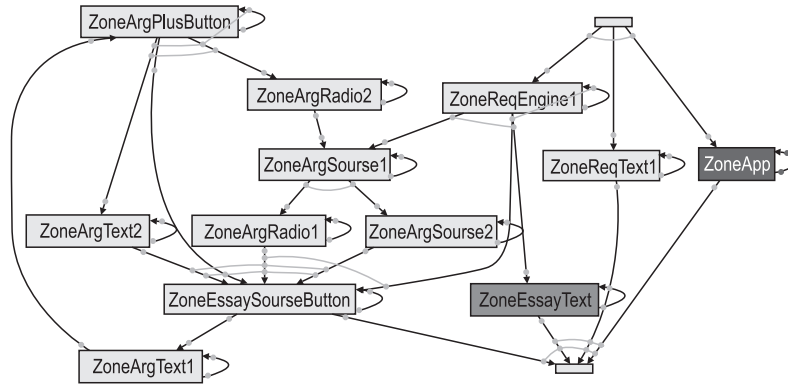
clearly see only the beginning and the end of the test (the two small rectangles with no captions), and all the other actions were happening rather haphazardly. The only insight one could get from scrutinising the graph could be into the amount of repetitions of the same activity. For example, the students left the app, came back, and left the app again 1636 times in total ("ZoneApp"). The round arrow pointing from and to the same box means that the activity paused and then proceeded. It may provide information about the density of using the internet (how frequently it was done in comparison to the other actions) if the number of students (38) is considered (see the Discussion). The same technique could be deployed for probing into other activities: the second most popular step is writing the essay ("ZoneEssayText") numbering 725 occurrences following one another (the semi-circular arrow shows that the same activity started when the previous step was finished). Yet, this graph is not easily digested when it comes to sequences of different events following one another. Hence, the next step to be taken is to study the C-net graph (Figure 5).

The boxes here indicate the same actions as in the previous graph, with the empty rectangles being the beginning (at the top) and end of the pattern (at the bottom). The light blue dots indicate that the activity could either go down the arrow to the next box, or up to the previous one. In other words, this infographic does not provide a clear linear representation of the steps. One thing that it does provide is understanding of combinations of activities, which are represented by the light blue lines (bindings) connecting the

arrows. These bindings mean that the majority of students completed all activities united by the binding in no particular order. To give an example, after writing the source of the first argument ("ZoneArgSource1"), the majority of the subjects either marked the argument as for/against ("ZoneArgRadio1") or added the next reference ("ZoneArgSource2"), and vice versa.

As with the low performers, the principal difference between the two graphs (Figures 4 and 5) lies in the focus of attention in the process. Figure 4 denotes the whole process flow (though it is impossible to see it clearly since students tend to repeat the same activities or come back to previous steps spontaneously), while Figure 5 was depicting the major staples of the process, the clusters of activities that the algorithm could extract from the most popular behaviour of the group. Figure 5 appears to be more informative than Figure 4 as it contains concrete steps that were commonly followed by other actions, while in Figure 4 no accurate information about sequences can be gained.

Figure 5. **C-net for high performers**



4. Discussion

Critical online reasoning is a complex latent construct that is difficult to measure. Doing multiple choice tests does not suit its nature: by its definition, it demands an open unconstrained online environment for subjects to show it. Thus, it is challenging to assess COR skills as it requires assessment of both the product and the process. Speaking of the former, there are successful attempts to design rubrics and implement evaluation. On the other hand, there is a paucity of research on the latter [McGrew et al., 2018].

However, investigating process patterns can provide an insight into the behaviour inherent in different levels of COR skills. Importantly, it may also be used as a source of evidence for further development and refinement of theoretical frameworks of the construct measured [Mislevy et al., 2003].

It has been shown in the previous section that being varied and tangled, general behaviour in different subject groups is barely possible to depict (Figures 2 and 4). There is no linear pattern, where one step always follows the other. The graph is rather complex and contains many concurrent actions. On the other hand, process mining software can effectively process fairly constrained behaviour, e.g. workflow or document flow. As was the case in Schmidt et al. (2020) research, graphic representations do not provide a clear understanding of a sequence of actions.

4.1. Similarities

As for the RQ, it is clear that the graphs of both low and high performing groups are similar in terms of their varied behaviour and non-linearity; it might be useful to glance at the most frequent combinations of activities for the utilitarian purpose of behaviour comparison (Figures 3 and 5). The low performers were creating a loop at the stage of googling and filling in the essay, query, and source form. The same was characteristic of the high-flyers. Another loop was occurring in the process of working with the argument and essay. However, only the high performers displayed a pattern of working on two arguments, which is remarkable as the task clearly stated to present at least one argument for and one against. Apart from these, the actions students tend to undertake are more or less of the same nature. This pattern discovery does not conflict with the previous research, indicating that there are no drastic differences between the set of most popular actions per se [Lai, 2011; McGrew et al., 2018; Weber et al., 2018]. Weber et al. [2018] provide a possible explanation, claiming that basic strategies are inherent to both groups whereas high achievers tend to implement basic tactics and refine them with more elaborate approaches to search.

As the literature review suggests, leaving the application in order to find an answer is deemed to be a successful strategy, while staying in the app and trying to come up with a solution by yourself is dismissed as a dead-end tactic [McGrew et al., 2019; Zlatkin-Troitschanskaia et al., 2020]. Surprisingly, in this research googling the subject matter online was not the case for the majority of high performers. While only 14% of low performers left the app as soon as they got the task, 13% of high performers did the same thing. The tendency displayed by the low performers coincides with other research results, whereas the high performers' general pattern does not [Shavelson et al., 2019; Zlatkin-Troitschanskaia et al., 2020] and should undergo further investigation. Modern theory also suggests that the behaviour of a "fact checker" (i.e. a high performing student) is right the opposite: leaving the page to search for the answer on the web [McGrew, 2018].

However, a possible explanation of such extraordinary behaviour might have been confusion amongst students, who expected the app to enable them to make a query. This could account for the majority of the whole pool (43% of low and 46% of high performers) starting off with filling in the request form. Yet, cognitive laboratories held prior to the main wave of testing did not indicate such confusion.

4.2. Differences

What does make the high performers stand out is the amount of time. On average, it took them 61% longer than the low performers to complete the task, which also implied that the former took more steps to submit the work. It is quite reasonable to say that if students spend more time on work, they are more likely to get better results, merely by filling in the form more carefully and attentively and not leaving any missing values, which was confirmed by the previous studies [Eichmann et al., 2020; Stadler et al., 2019; Tang et al., 2020; Ulitzsch et al., 2022].

Even more peculiar was the fact that the high performers switched to the internet on average three times as often as the low performing group did. The gulf between the amount of time the two groups invested in their work is very much in line with previous research, where the scholars showed that a longer internet search positively affects the outcomes [McGrew et al., 2019; Zlatkin-Troitschanskaia et al., 2020].

Another difference was in the number of arguments the two clusters submitted. As for the high performers, there were at least two, while for the low scoring students the number was one. Interestingly, there were low scoring students submitting two arguments (the required number); however, this behaviour was uncommon and thus filtered out during data preparation. This difference is also noticeable in other research [Zlatkin-Troitschanskaia et al., 2020], which revealed that students with prior beliefs about the issue tended to submit fewer arguments (importantly, the scholars also showed that prior beliefs were predictive of a lower COR level). Apparently, high performers do invest more time to look at the issue from multiple angles and later submit the fruits of their search.

A more structured way of comparing the two populations of low and high performers is presented in Table 3. There are six criteria by which the two groups were juxtaposed.

Table 3. **Comparison between low and high performers**

Unit	Low performers (63 students)	High performers (45 students)
Time in the app (min)	8	13
Time out of the app (min)	4	7

Unit	Low performers (63 students)		High performers (45 students)	
A mean number of actions	43		82	
A unique number of actions	13		20	
Switches between the app and the Internet	12		35	
The first action (% of students)	Fill in query form	0.43	Fill in query form	0.46
	Choose searching engine	0.30	Choose searching engine	0.24
	Leave app	0.14	Leave app	0.13
	Write essay	0.06	Write essay	0.07
	State the argument source	0.03	State the argument source	0.04
	Add a new query	0.01	Write an argument	0.02
			Add a new query	0.02

5. Limitations

An obvious limitation of the study lies in the fact that the analysis did not take into consideration steps happening outside the application, i.e. the students filled in the form only with the resources, links, and queries they deemed necessary, forgetting or deliberately omitting some steps of the process that happened on the Internet. The next step is to include into the analysis not only the log files in the application, but also those in the online environment (such as the amount of time spent on searching a particular web-site or the number of attempts taken to find the web-sites students refer to). The lack of evidence on how students spend their time online may have led to a less obvious division between patterns of the clusters. Advancing the instrument might enable one to see the difference between the two groups and register if there was any stark contrast between the behavioural patterns of high and low performers online.

On top of it, enhancing the techniques for and approaches to pattern extraction will provide an opportunity to use process data as a rich source of arguments in favour of test validity. There are more techniques for process analysis, which will be highlighted in the Conclusion and Future Research section.

Another limitation to be tackled by further research is the control of students' behaviour. In this study, the subjects were given the test and then instructed to complete it within a week. They were also said to take only 1 hour and 30 minutes for the testing. With the majority obeying the rules, some were taking much more time than it was allowed. What is more, their general attitude to the test was rather reluctant. It is advisable to carry out such research in class, observing students and limiting them in their time to submit works.

Also, one of the most important limitations is the fact that process-mining ignores individual differences between action sequences and is only focused on the cluster-based analyses, which was derived from the first part of the CT test (with standardised items) to form the clusters. The other ways of considering individual differences in the process-analysis results (e.g. by splintering the classes by the product score of the second part of the CT test) appears to be a very promising direction.

6. Conclusion and Future Research

Analysing graphs and diagrams provides researchers with some useful insights into common ways of fulfilling tasks. Digging deeper into the process of task completion is of paramount importance according to many pioneers in this area of scientific research as the process of solving the task can encompass either successful or unsuccessful strategies. [Wineburg, McGrew, 2017; McGrew et al., 2018; Weber et al., 2018; McGrew et al., 2019; Zlatkin-Troitschanskaia et al., 2021].

This study dealt with students' pattern analysis, where two clusters (low and high performers) were taken to draw the conclusions on the difference of the two. We could witness that the most prominent gulf is the one of time, namely, how much the high performers were ready to spare on the task in contrast with the low achievers. The former worked on average 61% longer. Another distinction was within the amount of actions generated by students while working with the app. As with the previous point, the high performers got almost twice as many actions as the low performers did on average. However, a stark difference was in the amount of time the students referred to googling. While solving the task, the high achievers googled extensively more than the low performers did. It resulted in generating three times more switches between the app and the internet. These differences are in line with previous research on COR, which extensively describes similar behaviour [Wineburg, McGrew, 2017; McGrew et al., 2018; Weber et al., 2018; McGrew et al., 2019; Zlatkin-Troitschanskaia et al., 2021].

However, the upshot of this work lies in the fact that there were no particular distinctions registered between general patterns students demonstrated on different COR levels. The graphs were barely readable, with numerous loops and steps that did not show the "bigger picture" of COR manifestation. What is more, it is yet unclear how to implement the mining techniques to assess the process. Nevertheless, it is desirable to be able to assess the process of task solution since it is embedded in the very definition of the construct of Critical Online Reasoning, where both the product and the process are essential. More advanced tools should be utilised in order to analyse the patterns. Potential for further research lies in the de-

ployment of Multidimensional Scaling, seq2seq autoencoders, Hidden Markov Models, and similar instruments.

Funding The article was prepared in the framework of a research grant funded by the Ministry of Science and Higher Education of the Russian Federation (grant ID: 075-15-2022-325).

References

Aalst van der W.M.P. (2016) *Process Mining: Data Science in Action*. Heidelberg: Springer.

Arpasat P., Premchaiswadi N., Porouhan P., Premchaiswadi W. (2021) Applying Process Mining to Analyze the Behavior of Learners in Online Courses. *International Journal of Information and Education Technology*, vol. 11, no 10, pp. 436–443. https://doi.org/10.18178/ijiet.2021.11.10.1547

Benevento E., Aloini D., Aalst W. (2022) How Can Interactive Process Discovery Address Data Quality Issues in Real Business Settings? Evidence from a Case Study in Healthcare. *Journal of Biomedical Informatics*, vol. 130, no 4, Article no 104083. http://dx.doi.org/10.1016/j.jbi.2022.104083

Care E., Kim H., Vista A., Anderson K. (2018) *Education System Alignment for 21st Century Skills: Focus on Assessment*. Washington, DC: Center for Universal Education at The Brookings Institution.

Dewey J. (1910) *How We Think*. Lexington, MA: D.C. Heath and Company. https://doi.org/10.1037/10903-000

Eichmann B., Greiff S., Naumann J., Brandhuber L., Goldhammer F. (2020) Exploring Behavioural Patterns during Complex Problem-Solving. *Journal of Computer Assisted Learning*, vol. 36, no 6, pp. 933–956. https://doi.org/10.1111/jcal.12451

Griffin P., McGaw B., Care E. (2012) The Changing Role of Education and Schools. *Assessment and Teaching of 21st Century Skills* (eds P. Griffin, B. McGaw, E. Care), Dordrecht, Germany: Springer Science+Business Media B.V., pp. 1–16. http://dx.doi.org/10.1007/978-94-007-2324-5_2

Hargittai E., Fullerton L., Menchen-Trevino E., Thomas, K.Y. (2010) Trust Online: Young Adults' Evaluation of Web Content. *International Journal of Communication*, vol. 4, no 1, pp. 468–494.

Kerr D., Andrews J.J., Mislevy R.J. (2016) The In-Task Assessment Framework for Behavioral Data. *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (eds A.A. Rupp, J.P. Leighton), Wiley-Blackwell, pp. 472–507. https://doi.org/10.1002/9781118956588.ch20

Kitchenham B., Charters S. (2007) *Guidelines for Performing Systematic Literature Reviews in Software Engineering. Version 2.4. EBSE Technical Report no EBSE-2007-01*. Keele, UK: Keele University.

Lai E.R. (2011) *Critical Thinking: A Literature Review Research Report*. London: Parsons.

Lightbown P.M., Spada N. (2021) *How Languages Are Learned*. Oxford: Oxford University.

Liu O.L., Rios J.A., Heilman M., Gerard L., Linn M.C. (2016) Validation of Automated Scoring of Science Assessments. *Journal of Research in Science Teaching*, vol. 53, no 2, pp. 215–233. http://dx.doi.org/10.1002/tea.21299

McGrew S., Breakstone J., Ortega T., Smith M., Wineburg S. (2018) Can Students Evaluate Online Sources? Learning from Assessments of Civic Online Reasoning. *Theory & Research in Social Education*, vol. 46, no 2, pp. 165–193. https://doi.org/10.1080/00933104.2017.1416320

McGrew S., Smith M., Breakstone J., Ortega T., Wineburg S. (2019) Improving University Students' Web Savvy: An Intervention Study. *British Journal of Educational Psychology,* vol. 89, no 3, pp. 485–500. https://doi.org/10.1111/bjep.12279

Mislevy R.J. (2018) *Sociocognitive Foundations of Educational Measurement*. New York, NY: Routledge.

Mislevy R., Almond R., Lukas J. (2003) *A Brief Introduction to Evidence-Centered Design. CSE Report no 632*. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Mislevy R.J. (2012) *Four Metaphors We Need to Understand Assessment.* Washington, DC: The Gordon Commission on the Future of Assessment in Education.

Molerov D., Zlatkin-Troitschanskaia O., Nagel M.T., Brückner S., Schmidt S., Shavelson R.J. (2020) Assessing University Students' Critical Online Reasoning Ability: A Conceptual and Assessment Framework with Preliminary Evidence. *Frontiers in Education*, vol. 5, December, Article no 577843. https://doi.org/10.3389/feduc.2020.577843

Padilla-García J.L., Benítez Baena I. (2014) Validity Evidence Based on Response Processes. *Psicothema*, vol. 26, no 1, pp. 136–144. http://dx.doi.org/10.7334/psicothema2013.259

Rozinat A., Aalst van der W.M.P. (2008) Conformance Checking of Processes Based on Monitoring Real Behavior. *Information Systems,* vol. 33, no 1, pp. 64–95. https://doi.org/10.1016/j.is.2007.07.001

Schmidt S., Zlatkin-Troitschanskaia O., Roeper J., Klose V., Weber M., Bültmann A.-K., Brückner S. (2020) Undergraduate Students' Critical Online Reasoning: Process Mining Analysis. *Frontiers in Psychology*, vol. 11, November, Article no 576273. https://doi: 10.3389/fpsyg.2020.576273

Shavelson R.J., Zlatkin-Troitschanskaia O., Beck K., Schmidt S., Marino J.P. (2019) "Assessment of University Students' Critical Thinking: Next Generation Performance Assessment". *International Journal of Testing*, vol. 19, no 4, pp. 337–362. https://doi.org/10.1080/15305058.2018.1543309

Stadler M., Fischer F., Greiff S. (2019) Taking a Closer Look: An Exploratory Analysis of Successful and Unsuccessful Strategy Use in Complex Problems. *Frontiers in Psychology*, vol. 10, May, Article no 777. https://doi.org/10.3389/fpsyg.2019.00777

Tang X., Wang Z., He Q., Liu J., Ying Z. (2020) Latent Feature Extraction for Process Data via Multidimensional Scaling. *Psychometrika*, vol. 85, no 2, pp. 378–397. https://doi.org/10.1007/s11336-020-09708-3

Tarasova K.V., Orel E.A. (2022) Measuring Students' Critical Thinking in Online Environment: Methodology, Conceptual Framework and Tasks Typology. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 187–212. https://doi.org/10.17323/1814-9545-2022-3-187-212

Ulitzsch E., He Q., Pohl S. (2021) Using Sequence Mining Techniques for Understanding Incorrect Behavioral Patterns on Interactive Tasks. *Journal of Educational and Behavioral Statistics,* vol. 47, no 1, pp. 3–35. https://doi.org/10.3102/10769986211010467

Ulitzsch E., Pohl S., Khorramdel L., Kroehne U., von Davier M. (2022) A Response-Time-Based Latent Response Mixture Model for Identifying and Modeling Careless and Insufficient Effort Responding in Survey Data. *Psychometrika,* vol. 87, no 2, pp. 593–619. https://doi.org/10.1007/s11336-022-09846-w

Verbeek H.M.W., Buijs J.C.A.M., van Dongen B.F., Aalst van der W.M.P. (2010) ProM 6: The Process Mining Toolkit. Proceedings of the *Business Process Management 2010 Demonstration Track (Hoboken NJ, USA, 2010, September 14–16),* pp. 34–39.

Weber H., Becker D., Hillmert S. (2019) Information-Seeking Behaviour and Academic Success in Higher Education: Which Search Strategies Matter for Grade

Differences among University Students and How Does This Relevance Differ by Field of Study? *Higher Education,* vol. 77, no 4, pp. 657–678. https://doi.org/10.1007/s10734-018-0296-4

Weber H., Hillmert S., Rott K.J. (2018) Can Digital Information Literacy among Undergraduates Be Improved? Evidence from an Experimental Study. *Teaching in High Education,* vol. 23, no 8, pp. 909–926. https://doi.org/10.1080/13562517.2018.1449740

Weijters A.J.M.M., Aalst van der W.M.P., Alves De Medeiros A.K. (2006) *Process Mining with the Heuristics Miner Algorithm. BETA Working Papers no 166*. Eindhoven: Technische Universiteit Eindhoven.

Wineburg S., McGrew S. (2017) *Lateral Reading: Reading Less and Learning More When Evaluating Digital Information. Stanford History Education Group Working Paper no 2017-A1*. http://dx.doi.org/10.2139/ssrn.3048994

Zieky M.J. (2014) An Introduction to the Use of Evidence-Centred Design in Test Development. *Psicología Educativa*, vol. 20, no 2, pp. 79–87. https://doi.org/10.1016/j.pse.2014.11.003

Zlatkin-Troitschanskaia O., Beck K., Fischer J., Braunheim D., Schmidt S., Shavelson R.J. (2020) The Role of Students' Beliefs When Critically Reasoning from Multiple Contradictory Sources of Information in Performance Assessments. *Frontiers in Psychology,* vol. 11, September, Article no 2192. http://dx.doi.org/10.3389/fpsyg.2020.02192

Zlatkin-Troitschanskaia O., Brückner S., Nagel M.-T., Bültmann A.-K., Fischer J., Schmidt S., Molerov D. (2021) Performance Assessment and Digital Training Framework for Young Professionals' Generic and Domain-Specific Online Reasoning in Law, Medicine, and Teacher Practice. *Journal of Supranational Policies of Education,* no 13, pp. 9–36. https://doi.org/10.15366/jospoe2021.13.001