PEM4PPM: A Cognitive Perspective on the Process of Process Mining

Elizaveta Sorokina[1] [0009-0000-4534-0912], Pnina Soffer[1] [0000-0003-4659-883X],

Irit Hadar[1] [0000-0002-4267-0235], Uri Leron[2],

Francesca Zerbato[3] [0000-0001-7797-4602] and Barbara Weber[3] [0000-0002-6004-4860]

[1] University of Haifa, City Campus, Hanamal St. 65, Haifa, Israel
esorokin@campus.haifa.ac.il, {spnina|hadari}@is.haifa.ac.il
[2] Technion - Israel Institute of Technology, Technion City, 3200003 Haifa, Israel
uril@technion.ac.il
[3] Institute of Computer Science, University of St. Gallen, Switzerland
{francesca.zerbato|barbara.weber}@unisg.ch

Abstract. During the last decades, process mining (PM) has matured and rapidly increased in its adoption. Making sense of data is a main part of the work of PM analysts, which involves cognitive processes. Recent work has leveraged behavioral data to explain these processes. Still, the process of process mining (PPM) is yet to be well understood and a theoretical foundation for explaining how these processes unfold is missing. This paper attempts to fill this gap by understanding how PPM data can be analyzed in a theory-guided manner and what insights can be gained from this analysis. To investigate these aspects, we analyzed verbal data and interaction traces obtained from analysis sessions with 29 participants performing a PM task. The analysis was based on the Predictive Processing (PP) theory and the derived Prediction Error Minimization (PEM) process, anchored in cognitive science. The results include (1) a theoretical adaptation of the PEM theory to the PPM context, (2) four strategies utilized by PM analysts, identified, and validated based on the adapted theory, and (3) an understanding of the differences in performance between analysts using different strategies and independence of the expertise level and the strategy choice.

Keywords: Process Mining, Predictive Processing, Prediction Error Minimization, Analysis Strategies, Mixed Methods

1 Introduction

Over the past decades, information systems penetrated all areas of people's lives and led to an incredible growth of produced event data in various branches of industry. This growth has led to the development of techniques to analyze such data, such as process mining, which stood out as a valuable discipline for extracting insights from event logs and supporting the discovery, monitoring, and improvement of real processes.

Until now, process mining research has mainly concentrated on the technical perspective, proposing new approaches and tool enhancements [1, 2]. Still, recent studies on the "process of process mining" (PPM) have highlighted the importance of the *individual* perspective and focusing on the individual analyst's perception and behavior [3].

The PPM includes different activities such as choosing the data sources, inspecting, cleaning, and preprocessing the data, selecting and implementing process mining algorithms, and interpreting and presenting the results. A main goal of individual analysts during the PPM is to make sense of data. The sense-making process entails an iterative cycle, where attention is focused according to a set goal, leading to the generation of predictions about the world, which are then tested and reconsidered for minimizing the prediction error, i.e., the discrepancy between the predicted and the actual input. These are complicated cognitive processes that may entail high cognitive load [4, 5]. Although recent works have attempted to leverage behavioral data to explain such processes, e.g., by eliciting patterns of observed behavior or analysts' challenges [6-10], these processes are not yet well understood and there is a lack of theory in the process mining (PM) field that can explain how these processes unfold.

This paper attempts to fill this gap by proposing a theory-guided analysis approach based on a theory from a neighboring field. More specifically, previously mentioned characteristics of the PPM point to a close correspondence with the Prediction Error Minimization (PEM) theory, which is a part of the Predictive Processing (PP) theory [11-13]. In this research, we adapted the PEM theory to the context of the PPM, yielding the PEM4PPM model. This model provides a theoretical lens for analyzing the PPM from a cognitive perspective and understanding the analysts' cognitive processes. We applied this model to analyze data recording analysts' interactions with PM tools as well as think-aloud data verbalizing their thinking processes while performing PM tasks during the analysis phase of the PPM. With our analysis, we aim to answer the following two research questions:

- RQ1: How can PPM data be analyzed in a theory-guided manner (specifically, based on the PEM theory)?
- RQ2: What insights can be gained from this analysis?

To address these questions, we employed a mixed-method (qualitative and quantitative) approach [14]. RQ1 was addressed in a qualitative manner. First, we made the theoretical adaptation of the PEM model for the PPM, which resulted in the first version of the PEM4PPM model. Second, data collection and analysis of students' data were performed with the goal to refine the model. As a result of this step, the PEM4PPM model was refined to include additional cognitive steps, and four analysis strategies were identified. Next, the model and the strategies were validated against data stemming from experienced analysts. In addition, an analysis protocol was established. Finally, we integrated our findings regarding the identified strategies from both datasets and formed hypotheses for RQ2. To answer RQ2, we applied a quantitative approach, focused on analyzing the influence of the previously discovered strategies on the analysts' performance and checking the association between the strategies and the level of expertise. The main contributions of this paper are (i) a theoretical adaptation of the PEM to the PPM context, (ii) four strategies utilized by PM analysts, identified and validated based on the adapted theory, and (iii) an understanding of the difference in performance between analysts using different strategies and independence of the expertise level and the strategy choice.

The remainder is organized as follows: Section 2 provides background concepts and related work. Section 3 presents the proposed PEM4PPM model. Section 4 describes the research method of this work and Section 5 reports on the findings. Section 6 discusses the findings and limitations of our work. Section 7 provides conclusions and outlines future research directions.

2 Related Work

2.1 Human Involvement in the PPM

Vom Brocke et al. [3] proposed a five-level framework for research on PM. Two of the levels, group and individual, focus on the humans involved in PM. While the group level characterizes the effect of PM on people's interaction and work mode, the individual level concentrates on people's perception and behavior. There are additional recently published papers taking an individual perspective and focusing on different aspects related to the behavior of PM analysts. These include analysts' patterns of behavior, goals, and strategies in the initial exploration phase [6], strategies applied by analysts during the analysis stage [7], the development of PM questions [8], the discovery of process improvement opportunities by exploring how analysts use PM [9], and challenges perceived by process analysts during different project phases [10].

This work also takes the individual perspective but extends the body of knowledge through a theory-guided approach. More specifically, this work attempts to unveil and explain the complicated cognitive processes that PM analysts follow to make sense of event data representing real-world behavior from a cognitive perspective [4, 5]. Such a theoretical cognitive lens is still missing.

2.2 The Prediction Error Minimization Theory

Our choice of a conceptual basis for analyzing the PPM from a cognitive perspective is based on the previously outlined similarities (cf. Section 1). PEM conceives the brain as a probabilistic inference system, which attempts to predict the input it receives by constructing models of the possible causes of this input. This system aims to minimize the prediction error, namely, the discrepancy between the predicted and the actual input. If the prediction error is small, then there may be no need to revise the model that gives rise to the prediction. If, on the other hand, the prediction error is large, then it is likely that the model fails to capture the causes of the inputs and, therefore, must be revised. In this case, there are two optional courses of action for reducing the prediction error. The first is to revise one's model of the world until the prediction error is satisfactorily diminished (termed ‘perceptual inference’) and the second is to change the world (e.g., by manipulations that yield additional inputs) so that it matches the model (termed ‘active inference’). Moreover, PEM suggests that predictive models are arranged in a hierarchy. Interpretation of sensory signals is at the lowest level of the hierarchy, while low-level features are grouped into objects at the higher levels. Even further up the hierarchy, these objects are grouped together as parts of larger scenes entailing multiple objects [11, 12]. It is important to note, that PEM is not only a formal apparatus which allows generating precise quantitative models, but it is also a general framework that can be applied in the service of different explanatory programs. Thus, this framework can be enriched with additional assumptions about a specific domain, e.g., PM [15].

3 The Proposed PEM4PPM Model Based on the PEM Theory

In this section, we present the PEM4PPM model, which we developed based on the PEM theory, by adaptation to the PPM as a part of answering RQ1 (cf. Section 4.1).

The PEM process involves receiving input, focusing attention (possibly based on a goal) [16], and then iteratively creating a model and respective predictions, testing predictions, minimizing errors and eventually, when the error is small enough, acting upon the prediction. The initial PEM4PPM model was created in a series of brainstorming sessions, where the bi-directional mapping between PEM steps and PM operations took place, considering what PM operations could accomplish each PEM step and what PEM step could correspond to each PM operation. The model is shown in Fig. 1, which illustrates a diagram of a goal-oriented PPM conceptualized in terms of the PEM theory. Each step or sequence of steps in the diagram can be skipped. Cognitive operations are color-coded according to the PEM steps to which they correspond, and example mining operations appear in italics below PPM steps where relevant.

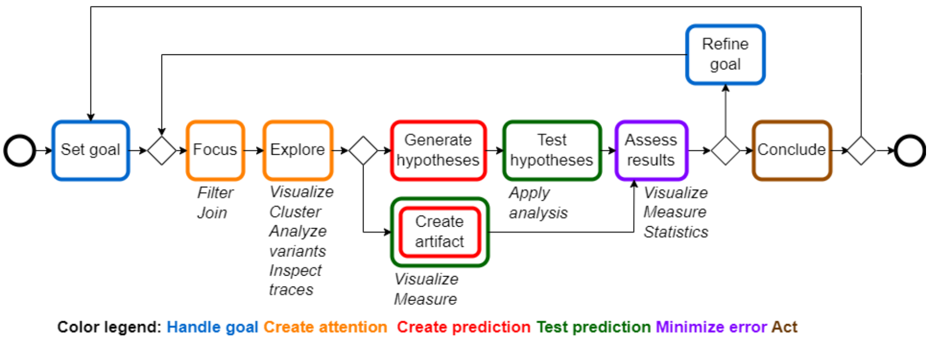

Fig. 1. The PEM4PPM model

According to the model, high-level business goals serve as a starting point for each PM endeavor. They can be later decomposed or refined into more specific goals. The refinement of goals can be done until the goal is concrete enough to be achieved by an available mining operation. Additional goals can be set or refined at any stage during this process. More concretely, when a goal is to be refined, a series of operations are involved. In the beginning, the relevant subset of the data needs to be filtered and organized. Next, this data needs to be explored to find behavior that may be of interest. These both are operations that enable focusing the attention on parts and aspects of the input data. Based on this, concrete hypotheses can be formed regarding the identified behavior, as predictions to be tested. Predictions can also be generated and tested by the creation of specific artifacts (e.g., creating a process model through discovery). Accordingly, available PM techniques are applied for testing hypotheses or creating artifacts. Then, the results are assessed against the goal or the hypothesis for minimizing the prediction error. There are two possible outcomes for this assessment: (1) conclusion – goal achieved; (2) the goal needs to be further refined and this can be repeated iteratively. Next, the results can point to a new goal to be set.

4 Research Methodology

This section presents the main steps of the research method followed in this paper.

4.1 Overview

The objective of this work was to gain an understanding of how to analyze PPM data in a theory-guided manner (based on the PEM model) (cf. RQ1) and to explore what insights can be gained from this analysis (cf. RQ2). For this purpose, a mixed-method research approach combining both qualitative and quantitative methods was applied [14]. Fig. 2 provides a high-level depiction of the method used for this research.

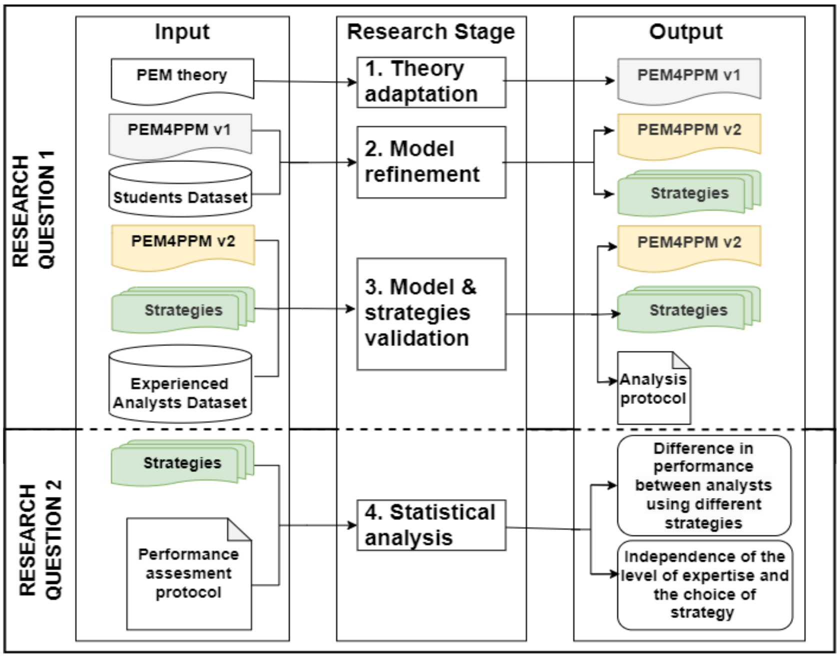

Fig. 2. Chronological visual representation of the mixed method process

To address RQ1 ("*How can PPM data be analyzed in a theory-guided manner (specifically, based on the PEM theory)?*") we applied a qualitative research approach [17]. Our aims were to adapt the PEM theory as a basis for analyzing the PPM and establish and validate an adapted PEM model and a derived analysis protocol.

First, during Stage 1, we adapted the PEM theory, creating a first version of the PEM4PPM model through brainstorming sessions (as described in Section 3).

Second, as a part of Stage 2, we refined the model using behavioral data. For this, we collected data from student participants (Students Dataset) and analyzed it. We decided to use students for model refinement since they were expected to have sufficient knowledge for performing a process mining analysis (all of them took the same "Process Mining" course). This stage resulted in the second version of the model, PEM4PPM v2, which included some additional, refined cognitive steps. Moreover, we used the model to analyze user behavior and derived strategies, as patterns of behavior, applied by the students during their work.

Finally, during Stage 3, we validated the PEM2PPM v2 model discovered in stage 2 and the identified strategies by analyzing the data that were collected from experienced analysts (Experienced Analysts Dataset). Experienced analysts were chosen for this stage since they differ in their expertise and experience from students and therefore could contribute to further validating, and possibly refining our findings. In addition, during this stage, we developed an analysis protocol describing the cognitive steps of the PEM4PPM v2 model.

To address RQ2 ("*What insights can be gained by analyzing the PPM on the basis of PEM model?*") a quantitative research approach was used. The strategies identified from analyzing the Students and Experienced Analysts Datasets in the context of RQ1, gave the direction for RQ2, i.e., to gain an understanding of whether the difference in performance between analysts using different strategies is significant and whether the level of expertise and the choice of strategy are independent of each other.

4.2 Data Collection and Settings

Overall, the study presented here comprises 29 participants: 16 students and 13 experienced analysts. More specifically, the Students Dataset was used for model refinement and for gaining initial insights (Stage 2), whereas the Experienced Analysts Dataset served for their validation (Stage 3). Despite having been collected separately, these two datasets are comparable since they were collected in the context of two studies of a similar design and set-up but different tasks and participants. More details about the two datasets are provided online: https://tinyurl.com/y28cxnzy, including demographic information, interview protocols and examples of experienced participants' coding for the entire duration of the session, showing the chain of evidence from statements to findings.

Collection of Data from Students. The data collection took place between June and July 2021. The study consisted of individual Zoom[1] video sessions with 16 participants,

[1] Zoom https://zoom.us/

all first- and second-degree students taking the "Process Mining" course provided by the University of Haifa[2]. All the students agreed to be observed and recorded via Zoom for the entire duration of the study. The study consisted of a process mining task (guided exploration) and a semi-structured interview. Before taking part in the study, students were asked to perform an initial exploration of the log guided by a few questions provided to them as a part of course assignments. This was done to help them avoid possible technical issues which may be related to various operations such as installing the tool, preprocessing, filtering, or importing/converting the log. In addition, these questions were designed to give students the opportunity to better understand the process, its activities and event attributes, and to familiarize with the tool before participating in the study. Our assumption was that this preliminary stage enabled participants to better concentrate on the provided question and guidelines during the data collection.

For the guided exploration, the students were asked to analyze the Road Traffic Fine Management (RTFM) event log, i.e., a real-life event log representing the process of managing fines for road traffic violations by the Italian police [18]. The log was chosen as it comes with extensive documentation about the process which allowed us to prepare the materials for the study [19]. The participants were asked to answer the following question related to the log "Describe two to three circumstances in which a penalty is added. Is the penalty addition always legitimate, i.e., do people that receive a penalty deserve it?". This question was designed to guide participants in exploring different scenarios but also go deeper into the data and investigate penalty addition. For their analysis, the participants could choose among different tools (Disco, ProM, PM4Py and bupaR), but all the participants chose Fluxicon Disco[3]. During the task, the participants were instructed to verbalize everything they were doing in a think-aloud manner [20]. In this way, we aimed to gain insights into their cognitive processes. The session was supervised by one author, who reminded participants to speak out in case of long sequences of actions happened without explanation. No feedback on their analysis strategies and outcomes was provided to the students during the session and they could conduct the analysis at their own pace.

After the participants completed the process mining task, we interviewed them to improve our understanding of analysis strategies and cognitive factors underlying them. The interviews were semi-structured [21] to allow expanding and gathering of information on relevant aspects emerging during the session.

For each student participant, we collected the recording of the session (screen recording of the analysis conducted in Disco and audio of the think-aloud), the log produced by Disco, i.e., a log that is produced automatically by the tool which records both the debug information and user's actions in the PM tool, and the audio recordings of the interviews. The session lasted up to one hour in most cases, with an average duration of 38 minutes. The interviews lasted between 8 and 13 minutes.

---

[2] The full study design, including participants' recruitment procedure, was approved by the Institutional Ethics Committee (Approval no. 238/21)

[3] Fluxicon Disco https://fluxicon.com/disco/

For this dataset, the signed participant consent, as approved by the Institutional Review Board (IRB) committee, does not allow for making this data publicly available in its raw form.

Collection of Data from Experienced Analysts. Thirteen experienced analysts served for validation, giving us the opportunity to validate the PEM4PPM v2 model, which we had refined based on the students' data. The data collection took place between May and July 2021, and involved analysts in the professional network of the authors who met the following requirements: (i) having analyzed at least two real-life event logs in the two years prior to the study and (ii) being knowledgeable of at least one among bupaR, Celonis, Disco, ProM and PM4Py, i.e., the PM tools available for the task. These two criteria allowed us to select participants familiar with the PM but having varying levels of experience and expertise. For this paper, we selected participants who used Disco to keep this set comparable to the Students Dataset and to reduce the dependency of the validation of the PEM4PPM v2 model and the strategies on the working mode imposed by the tool. The experienced analysts differ in their affiliation, job role and position, and tool knowledge. The data collection followed a similar design and the same materials used with the students: participants were asked to engage in a PM task while verbalizing their thoughts in a think-aloud manner. What differed was the scenario question we asked on the RTFM log: "Can you describe the three most prevalent circumstances in which the fines and the related expenses are not paid (in full) in this process?". Then, as done with the students, we prompted participants to look deeper into reasons for which fines are not paid. In this case, we did not ask participants to engage in an initial exploration phase, as the risks for technical issues and lack of experience with PM analysis were mitigated by the set-up (a remote desktop environment with the study materials and tools) and the participation requirements. Data collection was supervised by one author who prompted participants to speak in case of long silences, but, as for the students, no feedback on their analysis nor its outcome was given.

Finally, we conducted semi-structured interviews asking participants about (i) activities and artifacts, (ii) goals, (iii) strategies, and (iv) challenges of PM analysis. In detail, we asked them to reflect on the task they had just executed but also prompted them to share general anecdotes and work experiences with PM.

For each participant, we collected the same data as for the students: the screen and audio recording of the session, the log produced by Disco and the audio recordings of the interviews.

In this case, the sessions lasted between 20 and 90 minutes, with an average duration of 43. The interviews lasted between 23 and 43 minutes, and about 31 minutes on average. The data is collected as part of the ProMiSE project, for which we plan to share results and anonymized data with the research community upon project completion.

4.3 Data Preparation and Analysis

This section details how the collected data was analyzed to address RQ1 and RQ2 (Stages 2-4) leveraging the PEM4PPM model developed in Stage 1 (cf. Section 3).

Data Preparation: For the data preprocessing, we engaged in the following activities. First, (i) we transcribed the verbal utterances from the audio recordings of the session with the help of Vocalmatic[4] (Students Dataset) and MAXQDA[5] (Experienced Analysts Dataset), which allowed us to keep information about the timestamps of each utterance. Then, (ii) we watched all the screen recordings and reported next to each transcribed utterance a short description of non-verbal behavior observed from the videos. Examples of non-verbal behavior of interest for our analysis are pointing with the cursor to a specific process activity, reading documentation related to the task, filtering the data, etc. Then (iii) we merged and cleaned the logs from Disco removing debug information that did not relate to actions performed by participants in the tool and, thus, was not relevant to our research. As a final step, (iv) we merged and organized the data from (i), (ii) and (iii) in Excel, using the timestamps produced by the transcription tools and Disco to synchronize the different sources. In this way, we obtained a multimodal dataset [22] for our analysis, which included both verbal utterances and interaction traces of the participants with Disco.

Stage 2: To refine the PEM4PPM model we used the Students Dataset. This stage involved two activities: (i) coding the data according to the steps of the PEM4PPM v1 model to refine and tune the model and (ii) identifying different strategy types, characterizing, and grouping them into categories.

We utilized provisional coding, a first-cycle coding method, to code the Students Dataset [23]. The behavior and thinking process of each participant (as reflected by the verbal utterances and interaction traces) was analyzed individually. Since the steps of the PEM4PPM v1 model were not yet validated during this first analysis iteration, we analyzed the data described in the previous paragraph by combining both deductive and inductive methods. In detail, at the beginning of the analysis, a set of a priori codes, reflecting the steps of the PEM4PPM v1 model, was used. The coding was performed by assigning a code to each multimodal piece of data (think-aloud, screen recordings, and Disco logs), and using one or the other source to disambiguate in case the user cognitive step was not clear. The first author coded all the data in 3 iterations. At the end of each iteration, two other authors checked the codes independently to ensure consistency and assess the reliability of the process. Throughout the analysis, the authors revised and refined the codes collaboratively in several meetings. The final coding was determined through discussions among authors in an iterative consensus building process, similar to [24]. Since a one-to-one mapping of the data to the PEM4PPM v1 steps was not possible in all the cases, two additional steps (“Task understanding” and “Interpret data”) emerged as additional steps in the PPM process. Next, they were validated against the theoretical literature [25, 26] and we added them to our model, obtaining PEM4PPM v2.

Then, based on the PEM4PPM v2 model, high-level patterns were identified during the second cycle of coding. For this purpose, theoretical coding (sometimes also called “Selective Coding” or “Conceptual Coding”) [27] was utilized. It allowed us to group the first cycle summaries (steps of the PEM4PPM v2 model) into strategies applied by

4 Vocalmatic https://vocalmatic.com/
5 MAXQDA https://www.maxqda.com/

participants during their work. Analyzing commonalities and variabilities among these processes allowed us to identify the repetitive use of certain patterns and strategies that can be associated with different analyst profiles. These strategies were characterized by the presence or absence of the steps "Interpret data", "Generate hypotheses", or "Test hypotheses" in the verbal utterances, since other steps were present in most cases. These steps relate to cognitive operations of "creating prediction" and "testing". As a result of this analysis, we discovered four strategies associated with different analyst profiles. To investigate the possible reasons for applying specific strategies by students the interview data was examined.

Stage 3: To validate the PEM4PPM v2 model and the strategies discovered by analyzing the students' data, we coded and analyzed the data collected from experienced analysts. The coding and analysis of this dataset were performed following a similar approach to the analysis of the students' data described above. However, in this stage, we used the steps of the PEM4PPM v2 model in a deductive manner. The coding was performed by 2 authors in two iteration and, similarly to what was described for the students, we worked collaboratively in the team of authors to reach a consensus. The validation of the PEM4PPM v2 resulted in the analysis protocol describing the cognitive steps of the PEM4PPM v2 model, their explanation, example statement, and detection method. Table 1 presents a part of this protocol. The full analysis protocol can be found in the supplementary materials: https://tinyurl.com/y28cxnzy.

Cognitive step	**Explanation**	**Example statement**
Task understanding	Understanding the task/problem, understanding the data (e.g., attributes, model, activities), verifying with researcher/stakeholder information regarding the data or task.	"Wait, *the question is about* the fine or the additional penalty?" (S7)
Set/ Refine goal	Formulating a goal related to the question.	"So, the question is to identify the flows where the payment has not been done completely. That is why now I isolate the flows, the different flows we have..." (E26)
Focus	Concentrating on a specific part/activity of the map.	"So, I think I first want to *focus on the cases* in which there was the addition of a penalty, so I'm actually going to filter by this element [activity Add penalty]" (S3)
Explore	Learning about the process.	"He has a fine of 20, he has a fine of 22, he pays 22, then he had a penalty of 44. I don't understand why and then he paid another 35." (S8)
Interpret data	Explaining the process based on previously inspected behaviors, information about the process, the participant's previous	"Yeah, this is one third of the cases. Here it says 30% of the cases. We saw on the diagram that it was one third. Oh, yeah. 30% is one third. Um, yeah, one third of the citizens are good [laughing]. I have never driven in Italy, but it doesn't look like

	knowledge and experience, or their combination.	people drive following the rules. And I think that the data confirms..." (E3)
Create artifact	Producing deliverable objects driven by a goal.	"Hmm. And it will make a difference, probably. Is it possible for me *to make a screenshot* of this and put it aside?" (E3)
Generate hypotheses	Forming a hypothesis about previously identified behaviour or possible answers for the question.	"*I assume that* it can be a difference between Amount for example, paid and Amount counted." (E28)
Test hypotheses	Checking hypotheses using available PM techniques.	NA (this step can be identified only by actions in the PM tool performed to test previously generated hypotheses)
Assess results	Evaluating the findings against previously defined hypothesis or goal.	"There are actually quite a lot of cases where there was a payment in the dataset, there was no value for the Payment Amount." (E16)
Conclude	Providing a final answer to the question or its part.	"So, most of the fines are legitimate. The majority behaves in a way that is legitimate, but there are some cases, like I said, that either they paid and then received a penalty or they sent an appeal or received the appeal and did not receive an answer to it, but received a penalty. These are actually the illegitimate situations." (S5)

Table 1. Analysis protocol (partial) describing the cognitive steps of the PEM4PPM v2 model. For each quote, we write the student (S) and experienced (E) participant's ID in parenthesis.

Stage 4: Finally, aiming to address RQ2, we performed a combined analysis of both datasets focusing on the elicited strategies. Inspired by the analysis conducted to address RQ1, we formulated hypotheses regarding the difference in performance between analysts using different strategies and the independence of the level of expertise and the choice of strategy. In this stage, we tested such hypotheses. In order to assess the analysts' performance, we developed a gold standard: one author "graded" the answers provided by the participants, then two other authors validated the grades independently, and an iterative consensus building process took place (like the one applied for the coding). The answers were assessed using a grading scale from 0 to 100, where 0 indicates the lowest performance and 100 is the highest. The grades reflected the correctness of the analysts' answers to the questions of the task.

More details about the data analysis and findings, including the analysis protocol, the coding scheme with examples, the protocols for assessing the participants' performance and statistical outputs are available online: https://tinyurl.com/y28cxnzy.

5 Findings

In this section, we present the results of our analysis for RQ1 (cf. Section 3 and Section 5.1) and RQ2 (cf. Section 5.2).

5.1 Qualitative Analysis

Validated PEM4PPM model. During Stage 3 of the empirical study (cf. Section 4.1), we validated the PEM4PPM v2 model. The resulting model is shown in Fig. 3, additional steps when compared to the initial model version are highlighted with grey. The step “Task understanding” reflects situations in which participants spent time understanding the task at hand from the documentation available or asking clarification questions to the researcher supervising the data collection. Instead, “Interpret data” implies explaining the process based on previously observed behaviors, information about the process, the participant’s experience, or their combination.

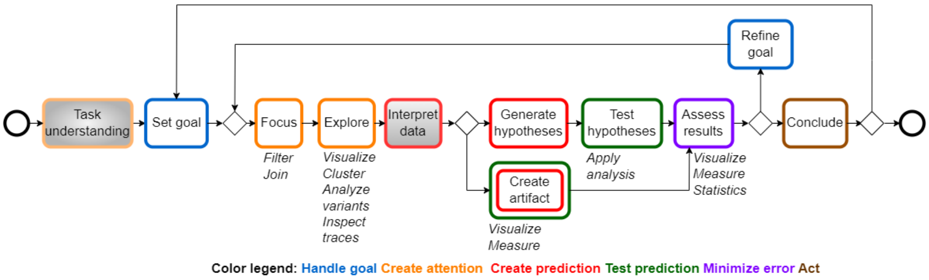

Fig. 3. The PEM4PPM model after validation (PEM4PPM v2)

Identifying and validating PM analysts' strategies. This section presents the four strategies applied by PM analysts. These strategies were identified using the Students Dataset during Stage 2 and validated with the data collected from experienced analysts in Stage 3. Here we present the strategies we identified, with sample participants’ quotes from interviews. For each strategy we indicate the count of students (S) and experienced analysts (E) using it in the form of #/29 (S = #, E = #):

1. *No indicated interpret data - No indicated hypothesis - No testing (NNN) strategy* which was observed for 2/29 (S = 2, E = 0) of the participants. The conclusions of the participants were based on their observations from the exploration stage. These participants spent the entire session understanding the data. “*I think I would try to use filters and not switch between cases and maps and try to draw conclusions through filters and this is also a way to work”.* (S12)
2. *With interpret data - No indicated hypothesis - No testing (WNN)* strategy which was observed for 12/29 (S = 8, E = 4) of the participants. Instead of basing their conclusions on the testing of hypotheses, these participants made them mainly based on exploration and data interpretation. "The *process was very new, for me, it was a really challenge to understand what's. It was a simple process of course. It was just finding and these things. But it was new for me, so it took a time for me to understand different steps of the process and the process was really strange in some ways.”* (E6)
3. With interpret data - With hypothesis - No testing (WWN) strategy which was observed for 5/29 (S = 2, E = 3) of the participants. Participants formulated hypotheses but for some reason (e.g., insufficient tool knowledge (S8, S16, E28) or question

understanding (S16), were reluctant to follow a trial-and-error approach while being observed (E11)) did not test them. *"Indirectly, knowing that you were looking at the screen while doing this, I think not consciously or unconsciously, I think I was a little bit reluctant on the trial and error."* (E11)

4. *With interpret data - With hypothesis - With testing (WWW)* strategy which was observed for 10/29 (S = 4, E = 6) of the participants. The WWW strategy reflects the PEM4PPM v2 model by containing all its steps. "So, *it was not like you do the hypothesis and then you check it is more like you do a preliminary hypothesis. You check that the hypothesis make sense in a secondary step once the data is already been explored and then you do, you check the hypothesis after that always related with the questions."* (E7)

It is important to note that "No indicated interpret data" or "No indicated hypothesis" means that these cognitive steps were not explicitly expressed during the session by the participants. Although they could be performed implicitly, they are clearly not emphasized. Looking through the data after its classification according to the strategies, we noticed that experienced analysts tend to apply the WWW strategy (46%) more than students (25%).

Moreover, we observed that the answers of participants utilizing the WWW strategy are of higher quality (in terms of correctness and accuracy) than those of participants using other strategies. The students' mean grade was 51, while the experienced analysts' mean grade was 59. These assessments, together with the defined strategies, formed a basis for formulating four hypotheses:

- H1a: analysts (students and experienced) who use the WWW strategy have <u>higher</u> performance (grades) than analysts using other strategies.
- H1b: students who use the WWW strategy have <u>higher</u> performance (grades) than students using other strategies.
- H1c: experienced analysts who use the WWW strategy have <u>higher</u> performance (grades) than experienced analysts using other strategies.
- H1d: the level of expertise (student vs experienced analyst) and the choice of strategy (WWW vs any other strategy) <u>are not independent.</u>

5.2 Quantitative Analysis

The hypotheses formulated based on the findings of RQ1 aimed to investigate the differences in performance between analysts using different strategies and check whether the level of expertise and the choice of strategy are independent of each other. For the analysis, the data regarding the strategy applied by each participant and their performance was organized and exported to an IBM SPSS version 23 environment for statistical analysis.

In order to test hypotheses a-c, we conducted the Mann-Whitney test (also known as the Wilcoxon test for independent samples) [28]. This test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed.

The conclusions resulting from testing hypotheses a-c (Table 2) is that the grades of analysts (in general, and students and experienced analysts separately) using the WWW strategy are significantly higher than the grades of those using other strategies.

Hypothesis	**N**	**U**	**Z**	**Calculated effect size (r)**	**One-tailed p-value**	**Decision for H0 (reject H0 if p-value < α (0.05))**
a (S and E)	29	13.5	-3.762	-0.699	.0000185	rejected
b (S)	16	2.5	-2.688	-0.672	.002	rejected
c (E)	13	6	-2.161	-0.599	.0175	rejected

Table 2. Results of Mann-Whitney test; students (S), experienced analysts (E).

To check whether selecting the WWW over any other strategy is independent of being a student or an experienced analyst (hypothesis d), we conducted the Chi-Squared Test of Independence of Categorical Variables [29]. It is a statistical test used to check whether some categorical variables are statistically independent in a population.

The conclusion from testing hypothesis d is that H0 cannot be rejected ($\chi\,(1)^2= 1.421$, p-value = .233). Therefore, we concluded that the choice of the WWW strategy over another does not depend on being a student or an experienced analyst.

6 Discussion

In this section, we review our findings, link them to related work, and discuss their impact on future research. Our work proposes the PEM4PPM model, a theoretical foundation for explaining the cognitive processes of individual analysts during the PPM, derived from theory and refined and validated with user behavior data. This approach complements fully data-driven work, e.g., [6-10], and extends the body of knowledge on the individual perspective with a theoretical lens.

In addition, based on the PEM4PPM model we identified four strategies applied by PM analysts. While some analysts performed all operations (WWW strategy), other analysts did not "generate hypothesis" and "test hypothesis" (NNN and WNN strategies). A possible explanation for not investigating the data more in-depth may be that analysts pursuing the NNN and WNN strategies were still in the stage where they collected evidence in the hope to suggest a hypothesis, rather than in the stage of testing a hypothesis they had in mind [30]. The development of strategies aiming to improve performance by realizing all their steps, usually in a specific order, is well renowned in the problem-solving and information-processing field [31, 32]. Analysts' strategies were also the focus of [7]. Unlike this work they organized them by analysis goals (understand, plan, analyze, and evaluate) and based their analysis on post-session interviews, rather than on the participants' real-time think-aloud.

The established analysis protocol together with the developed and validated PEM4PPM model can be used for future research. Moreover, this has implications for tool development or improvement allowing to consider the individual and his/her analysis processes. Since identified strategies differ in terms of performance this can give rise to the development of process guidance for PM analysts. These findings also can have implications for teaching students how to approach process mining. Beyond the expected contributions to the field of process mining, the outcomes of this research may be possible to generalize and adapt for additional data science tasks, which also entail extracting knowledge from data.

Limitations. The findings of this study should be considered in the light of some limitations. First, although participants were constantly encouraged to verbalize their thoughts in a think-aloud manner, it is possible that some thoughts might have remained concealed. We mitigated this risk by triangulating the verbal utterances with the log data. In the future, adopting additional techniques, e.g., eye tracking or EEG [33], might help to fill this gap, since additional fine-grained information about where the analyst focused their attention might be collected. Second, although the data coding was done carefully and iteratively by building consensus among researchers, it is still based on subjective assessment and could be done differently. Yet, we note that this is a common procedure [24], and that such an iterative process, where coding changes along the iterations, intercoder reliability is usually not measured. Third, while the findings reported here have been established, specifically with statistically significant results for the quantitative analysis, the setting of the study is limited, using one dataset and one process mining tool. For addressing our research questions, of how to analyze PPM data and what insights can be gained, this is sufficient. Yet, to draw general conclusions about PPM, additional studies are still needed. At last, one important consideration to make is the confidentiality of the collected data, which, as clarified in Section 3, cannot be fully shared. While we acknowledge that this may limit the replicability of the study, the inclusion of representative examples adds transparency and enables the reader to follow the chain of evidence from the raw data to the presented results.

7 Conclusions

In this paper, we have presented the findings of an empirical study investigating how PPM data can be analyzed in a theory-guided manner (specifically, based on the PEM theory) and what insights can be gained from this analysis. We adapted the PEM theory to the context of the PPM, yielding the PEM4PPM model. This model provided a theoretical lens for analyzing the PPM from a cognitive perspective and understanding the analysts' cognitive processes. Applying the PEM4PPM model for the analysis of the verbal utterances and interaction traces obtained from sessions with 29 participants we identified four strategies applied by PM analysts. We characterized these strategies by the sets of cognitive operations involved. Moreover, we investigated whether the difference in performance between analysts using different strategies is significant and whether the level of expertise and the choice of strategy are independent. We found that: (i) the grades of analysts (in general, and students and experienced analysts separately) using the WWW strategy are significantly higher than the grades of those using other strategies; (ii) the level of expertise (student or experienced analyst) and the choice of strategy are independent.

Future Work. We established PEM4PPM as a basis for analyzing PPM data and showed that it can yield benefits. Having established this, the generalizability of the findings may be improved by conducting additional studies with more participants and in different settings e.g., different PM tools and tasks using different event logs. In addition, a better understanding of the thinking processes of PM analysts may be gained by collecting data with technologies such as eye tracking or EEG. Another possible avenue for future research is the development of a support tool, designed in accordance with the PEM4PPM model, to assist PM analysts throughout the PPM.

Acknowledgment. This work was partially funded by the Israel Science Foundation under grant agreement 2005/21 and the Swiss National Science Foundation (SNSF) under Grant No.: 200021 197032.

References

1. R'bigui, H., Cho, C.: The state-of-the-art of business process mining challenges. International Journal of Business Process Integration and Management. 8, 285 (2017). https://doi.org/10.1504/ijbpim.2017.10009731.
2. Tiwari, A., Turner, C.J., Majeed, B.: A review of business process mining: state-of-the-art and future trends. Business Process Management Journal. 14, 5–22 (2008). https://doi.org/10.1108/14637150810849373.
3. vom Brocke, J., Jans, M., Mendling, J., Reijers, H.A.: A Five-Level Framework for Research on Process Mining. Business & Information Systems Engineering. 63, 483–490 (2021). https://doi.org/10.1007/s12599-021-00718-8.
4. Sweller, J.: Cognitive Load During Problem Solving: Effects on Learning. Cognitive Science. 12, 257–285 (1988). https://doi.org/10.1207/s15516709cog1202_4.
5. Jans, M., Soffer, P., Jouck, T.: Building a valuable event log for process mining: an experimental exploration of a guided process. Enterprise Information Systems. 13, 601–630 (2019). https://doi.org/10.1080/17517575.2019.1587788.
6. Zerbato, F., Soffer, P., Weber, B.: Initial Insights into Exploratory Process Mining Practices. Lecture Notes in Business Information Processing. 145–161 (2021). https://doi.org/10.1007/978-3-030-85440-9_9.
7. Zerbato, F., Soffer, P., Weber, B.: Process Mining Practices: Evidence from Interviews. Lecture Notes in Computer Science. 268–285 (2022). https://doi.org/10.1007/978-3-031-16103-2_19.
8. Zerbato, F., Koorn, J.J., Beerepoot, I., Weber, B., Reijers, H.A.: On the Origin of Questions in Process Mining Projects. Lecture Notes in Computer Science. 165–181 (2022). https://doi.org/10.1007/978-3-031-17604-3_10.
9. Kubrak, K., Milani, F., Nolte, A.: Process Mining for Process Improvement - An Evaluation of Analysis Practices. Research Challenges in Information Science. 214–230 (2022). https://doi.org/10.1007/978-3-031-05760-1_13.
10. Zimmermann, L., Zerbato, F., Weber, B.: Process Mining Challenges Perceived by Analysts: An Interview Study. Enterprise, Business-Process and Information Systems Modeling. 3–17 (2022). https://doi.org/10.1007/978-3-031-07475-2_1.
11. Hohwy, J.: The Self-Evidencing Brain. Noûs. 50, 259–285 (2014). https://doi.org/10.1111/nous.12062.

12. Clark, A.: Busting Out: Predictive Brains, Embodied Minds, and the Puzzle of the Evidentiary Veil. Noûs. 51, 727–753 (2016). https://doi.org/10.1111/nous.12140.
13. de Bruin, L., Michael, J.: Prediction error minimization: Implications for Embodied Cognition and the Extended Mind Hypothesis. Brain and Cognition. 112, 58–63 (2017). https://doi.org/10.1016/j.bandc.2016.01.009.
14. Venkatesh, V., Brown, S.A., Bala, H.: Bridging the Qualitative-Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems. MIS Quarterly. 37, 21–54 (2013). https://doi.org/10.25300/misq/2013/37.1.02.
15. de Bruin, L., Michael, J.: Prediction Error Minimization as a Framework for Social Cognition Research. Erkenntnis. 86, 1–20 (2018). https://doi.org/10.1007/s10670-018-0090-9.
16. Williams, D.: Predictive Processing and the Representation Wars. Minds and Machines. 28, 141–172 (2017). https://doi.org/10.1007/s11023-017-9441-6.
17. Miles, M.B., Huberman, A.M., Saldana, J.: Qualitative Data Analysis. (2019).
18. de Leoni, Mannhardt, and F. Mannhardt. "Road traffic fine management process." Eindhoven University of Technology. Dataset (2015).
19. Mannhardt, F.: Multi-perspective process mining. Ph.D. thesis, Technische Universiteit Eindhoven (2018).
20. Denzin, N.K., Ericsson, K.A., Simon, H.A.: Protocol Analysis: Verbal Reports as Data. Contemporary Sociology. 14, 125 (1985). https://doi.org/10.2307/2070501.
21. Myers, M.D., Newman, M.: The qualitative interview in IS research: Examining the craft. Information and Organization. 17, 2–26 (2007). https://doi.org/10.1016/j.infoandorg.2006.11.001.
22. Giannakos, M.N., Sharma, K., Pappas, I.O., Kostakos, V., Velloso, E.: Multimodal data as a means to understand the learning experience. International Journal of Information Management. 48, 108–119 (2019). https://doi.org/10.1016/j.ijinfomgt.2019.02.003.
23. Saldaña, J.: The coding manual for qualitative researchers. Sage (2021).
24. Recker, J., Safrudin, N., Rosemann, M.: How novices design business processes. Information Systems. 37, 557–573 (2012). https://doi.org/10.1016/j.is.2011.07.001.
25. Westbrook, L.: Mental models: a theoretical overview and preliminary study. Journal of Information Science. 32, 563–579 (2006). https://doi.org/10.1177/0165551506068134.
26. Staggers, N., Norcio, A.F.: Mental models: concepts for human-computer interaction research. International Journal of Man-Machine Studies. 38, 587–605 (1993). https://doi.org/10.1006/imms.1993.1028.
27. Strauss, A., Corbin, J.: Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. (1998). https://doi.org/10.1604/9780803959392.
28. Nachar, N.: The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. Tutorials in Quantitative Methods for Psychology. 4, 13–20 (2008). https://doi.org/10.20982/tqmp.04.1.p013.
29. Mendenhall, W., Beaver, B.M., Beaver, R.J.: Mendenhall's Brief Introduction to Probability and Statistics. Brooks/Cole (2001). https://doi.org/10.1604/9780534396091.
30. Baron, J.: Thinking and Deciding. (2006). https://doi.org/10.1017/cbo9780511840265.
31. Wallace, A.F.C.: Plans and the Structure of Behavior. George A. Miller, Eugene Galanter, Karl H. Pribram. American Anthropologist. 62, 1065–1067 (1960). https://doi.org/10.1525/aa.1960.62.6.02a00190.
32. Polya, G.: How to Solve It: A New Aspect of Mathematical Method. (2004). https://doi.org/10.1604/9780691119663.
33. Dehaene, S., Lau, H., Kouider, S.: What Is Consciousness, and Could Machines Have It? Robotics, AI, and Humanity. 43–56 (2021). https://doi.org/10.1007/978-3-030-54173-6_4.