

RESEARCH

Open Access



Beyond process model complexity: a multi-granular investigation across time and space based on eye-tracking

Thierry Sorg^{1*} , Amine Abbad-Andaloussi¹ , Ekkart Kindler² and Barbara Weber¹

*Correspondence:

Thierry Sorg
thierry.sorg@unisg.ch

¹Institute of Computer Science,
University of St. Gallen, St Gallen,
Switzerland

²DTU Compute, Technical University
of Denmark, Kongens Lyngby,
Denmark

Abstract

Process models support analysis, design, implementation, and operation of information systems and must therefore remain understandable for diverse process stakeholders. Numerous metrics estimate model complexity, yet their ability to predict the mental effort (i.e., cognitive load) users actually experience is still unclear. This study addresses that gap with a controlled eye-tracking experiment involving 27 participants and process models that systematically vary in essential complexity (logic) and accidental complexity (layout). Some of these models combine simple and complex parts. Task complexity is explicitly manipulated so that questions target regions of differing complexity. A coarse-grained level analysis tests how model and task complexity relate to cognitive load. Then, a fine-grained level analysis segments eye-tracking data in space and time while investigating users' visual behavior and cognitive load during model comprehension. Results show that complexity metrics align with cognitive load when the essential and accidental complexity of process models are well captured, and that the task characteristics affect this relationship. Moreover, they show that distinct phases emerge during process model comprehension and that non-task-relevant regions in process models contribute to cognitive load in a non-uniform way (especially early in a task). These findings strengthen measurement practice in Business Process Management (BPM) by establishing an empirical correspondence between a comprehensive metric suite and multimodal cognitive load indicators and point to adaptive modeling tools that help focus attention on task-relevant parts of process models.

Keywords Process model, Complexity metric, Cognitive load, Visual behavior, Difficulty, Eye tracking

Introduction

Process models anchor every phase of an information system lifecycle (Dumas et al. 2013). During analysis and design, they externalize requirements; during implementation, they guide configuration; and during operation, they enable monitoring and optimization (Dumas et al. 2013). Because these artifacts must be read by software engineers, domain experts, and managers alike, they are typically expressed in visual

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

notation, ranging from Goldstine-von Neumann flowcharts of the 1940s (Goldstine and Neumann 1947) to contemporary languages such as the Unified Modeling Language (UML)¹ and Business Process Model and Notation (BPMN)². Visual notations thus form part of what Moody calls the “language of software engineering” (Moody 2009) and, like any language, must remain cognitively manageable to fulfill their communicative purpose (Dumas et al. 2013).

To anticipate comprehension problems early, researchers have proposed dozens of complexity metrics for process models (Cheng 2008; 2012; Mendling 2008). However, while prior studies showed that complexity metrics correlate with errors in process models (Mendling 2008), it remains unclear to what extent they can predict the mental effort (i.e., cognitive load) actually experienced by users. Empirical studies in neighboring domains report conflicting evidence (Couceiro et al. 2019; Peitek 2022; Hao et al. 2023), suggesting that other aspects may affect the mapping from ‘objective’ complexity to ‘felt’ difficulty.

First, it is crucial to well capture and separate essential complexity that is inherent in a model’s control-flow logic, from accidental complexity, which arises from the representation, structure, or layout of the model. Second, cognitive load varies along two axes. The task-context axis reflects which part of the model the current task activates, while the temporal-dynamics axis captures how mental effort (i.e., cognitive load) evolves over the task’s timeline—for example, from broad search for task-relevant information to focused inference of how two process model activities relate to each other. Progress therefore hinges on untangling these complexity sources and observing user behavior at both coarse and fine temporal resolutions.

Building on and extending the eye-tracking study of Sorg et al. 2025, we conducted an experiment in which 27 participants performed comprehension tasks on process models that systematically vary along the essential and accidental complexity dimensions. Our study gathers subjective, performance, physiological, and behavioral indicators of cognitive load and analyzes them at two complementary levels (coarse vs. fine-grained).

Although our previous work (Sorg et al. 2025) investigated the relationship between process model complexity and cognitive load, it used only basic process models that are either simple or complex as a whole and reported primarily qualitative coarse-grained findings. We extend that line of research in three ways. First, we introduce a second set of *advanced* models that embeds low and high complexity parts within the same diagram, allowing us to test whether the parts irrelevant to the task affect cognitive load. Second, we explicitly manipulate task complexity. That is to say, we manipulate the tasks so that each participant solves comprehension tasks targeting regions of distinct complexity; this enables a direct test of whether task characteristics affect the alignment between complexity metrics and measures of cognitive load. Third, we bring fine-grained analysis to bear: We segment gaze data reflecting users’ visual behaviors (captured by eye-tracking) spatially (task-relevant vs. non-task-relevant regions) and temporally (discovery vs. post-discovery phases), and examine the resulting measures using descriptive and inferential statistics.

We formulate our research questions as follows:

¹<https://www.uml.org/>

²<https://www.bpmn.org/>

- Coarse-grained level research questions:
 - **RQ1:** Does process model complexity, as captured by existing complexity metrics, align with measures of users' cognitive load?
 - **RQ2:** Does the complexity of tasks affect the alignment between existing complexity metrics and measures of users' cognitive load?
- Fine-grained level research questions:
 - **RQ3:** How do users' visual behaviors evolve over time during comprehension tasks on process models of varying complexity?
 - **RQ4:** Does the complexity of the non-task-relevant parts of process models affect measures of users' cognitive load during comprehension tasks?
 - **RQ5:** How does the complexity of the non-task-relevant parts of process models affect users' cognitive load during different phases of comprehension tasks?

Contributions

This study extends prior work on process model comprehension and cognitive load by combining a multi-granular experimental design with fine-grained eye-tracking analysis. Our main contributions are as follows:

- **Established correspondence between complexity metrics and estimates of cognitive load.** We demonstrated this relationship through testing whether existing process model complexity metrics relate to cognitive load measures collected from users performing comprehension tasks. Building on prior studies that link these metrics to model understandability during cognitive tasks Mendling 2007; Bernstein and Soffer 2015; Burattin et al. 2017, we selected metrics that quantify either essential or accidental complexity and examined how they relate to intrinsic and extraneous cognitive load.
- **Influence of task characteristics on cognitive load.** We analyze how task complexity affects the relationship between process model complexity and cognitive load. By comparing tasks that address parts of models with varying complexity, our results reveal that task characteristics play a significant role in this relationship.
- **Fine-grained analysis of visual behavior.** We identify and characterize distinct phases of visual exploration ('discovery' vs. 'post-discovery') and show how they correspond to different patterns of cognitive load. This contributes to a deeper understanding of the temporal dynamics of model comprehension and links observed gaze behavior to sensemaking processes (Chen et al. 2025).

These insights advance the measurement theory in Business Process Management (BPM) and inform the design of adaptive modeling tools.

The remainder of the paper is organized as follows. Section [Background and Related Work](#) reviews the background and related work; Section [Complexity Metrics and Cognitive Load](#) presents the selected process model metrics employed in our analysis; Section [Research Method](#) details our research method, experimental design, and data analysis pipeline; Section [Findings](#) presents the results; Section [Discussion](#) discusses the results, implications and limitations; Section [Conclusion and future work](#) concludes the paper.

Background and related work

This section establishes the theoretical foundation for this research. Section [Complexity of Process Models and Comprehension Tasks](#) provides an overview of the research on complexity of process models and associated comprehension tasks. Section [Cognitive Load Theory, Measurements and Granularity](#) discusses cognitive load theory, measurement techniques and the level of granularity on which they are applied. Section [Visual Behavior Analysis](#) explores visual behavior analysis methods.

Complexity of process models and comprehension tasks

Research documents multiple metrics to measure process model complexity, originating from domains including software engineering, graph theory, and information theory (Mendling 2007; Sánchez-González et al. 2010). Following prior work (Antinyan 2020; Abbad-Andaloussi 2023), we classify these metrics into two categories, capturing essential and accidental complexity, respectively.

Complexity of process models

Essential complexity metrics concentrate on inherent characteristics of the logic that encodes the process specifications. For instance, (Mendling 2007) developed a comprehensive set of metrics encompassing dimensions of essential complexity, including *size*, *density*, *partitionability*, *connector interplay*, *cyclicity*, and *concurrency*. In contrast, accidental complexity metrics address aspects related to model presentation and layout. The measures introduced by (Bernstein and Soffer 2015) (specifically *edge style*, *crossing edges*, *edge angles*, and *symmetry in blocks*) along with the *flow consistency* metric from (Burattin et al. 2017) allow one to assess different aspects of accidental complexity. Our investigation employs these metrics to assess both essential and accidental complexities and their impact on cognitive load (cf. Section [Background and Related Work](#)).

Complexity associated with comprehension tasks

Previous studies investigating process model complexity have relied primarily on comprehension metrics—time and accuracy (Reijers and Mendling 2011; Mendling et al. 2012; Petrusel et al. 2017)—together with error counts (Mendling 2007) and self-reported difficulty (Petrusel et al. 2017). To our knowledge, the only work that explicitly applies the cognitive load theory is (Abbad-Andaloussi et al. 2023), who examined declarative (constraint-based) models. Instead, we focus on imperative (flow-based) models, which differ in their structural properties and are widely used in practice (Pichler et al. 2011; Fahland et al. 2009). In addition, we enrich previous approaches (e.g. (Mendling 2007; Reijers and Mendling 2011; Mendling et al. 2012; Petrusel et al. 2017)) by combining subjective, performance, physiological, and behavioral measures, allowing a multimodal analysis of user cognitive load at both coarse- and fine-grained levels of granularity.

While complexity metrics typically quantify the overall complexity of process models, comprehension tasks often focus on specific model parts rather than the model as a whole (Petrusel and Mendling 2013). These fragments, known as task-relevant parts, are thought to drive the mental effort exerted by model readers and account for much of the perceived difficulty (Petrusel et al. 2017). Therefore, when assessing process model

complexity, it is crucial to account for the specific tasks the models are intended to support (Malinova and Mendling 2021).

Prior research has emphasized the critical role of tasks in shaping process model comprehension (Malinova and Mendling 2021) and demonstrated that specific task-related characteristics can significantly impact users' understanding of process models (Vessey and Galletta 1991, Ritchi et al. 2019, Schreiber et al. 2023, 2024). These characteristics distinguish, for instance, between schema tasks, which rely solely on information within the model, and non-schema tasks, which require external domain knowledge (Ritchi et al. 2019); symbolic tasks, which focus on individual diagram elements, and spatial tasks, which emphasize relationships among elements (Vessey and Galletta 1991); as well as local tasks, which concern elements present within the same view, and global tasks, which target elements that span across multiple views (Schreiber et al. 2023, 2024).

In this study, we extend this characterization by introducing a distinction based on model complexity. Herein, we differentiate between tasks that focus on process model parts with a low level of complexity, referred to as simple tasks, and those that target parts with a higher level of complexity, referred to as complex tasks.

Cognitive load theory, measurements and granularity

Cognitive Load Theory (CLT) postulates that human working memory possesses finite capacity (Sweller et al. 2011; Chen et al. 2016). When exceeding these limits, cognitive overload occurs, characterized by impaired performance and elevated error rates (Chen et al. 2016). CLT differentiates three cognitive load types: intrinsic, extraneous, and germane (Sweller 2010). Intrinsic load derives from the inherent complexity of the subject material. For process models, this corresponds to their essential complexity (see Sect. [Introduction](#)) (Antinyan 2020; Abbad-Andaloussi et al. 2023). Extraneous load emerges from poor quality of information presentation, which is related to accidental complexity in process models (see Sect. [Introduction](#)) (Antinyan 2020; Abbad-Andaloussi et al. 2023). Germane load, finally, concerns the construction of mental schemes for effective information processing using existing knowledge (Sweller 2010). Our investigation analyzes how essential and accidental complexities influence intrinsic and extraneous loads, respectively. In addition, our within-subject experiment controls germane load factors and keeps them constant to limit the impact of this load (see Sect. [Experimental Procedure](#)).

Different measures are known for their ability to capture cognitive load (Chen et al. 2016). The following paragraphs present a set of these measures, which we are using in our study (cf. Sect. [Study Design](#)).

Subjective measures

These measures evaluate task difficulty through self-reported assessments. A typical example is the 5-point Likert scale for perceived difficulty (ranging from *very easy*, indicating low load, to *very difficult*, indicating high load) (Abbad-Andaloussi et al. 2023; Schreiber et al. 2023).

Performance measures

These measures assess the performance of the user regarding the execution of the task (Chen et al. 2016). For comprehension tasks, typical measures incorporate accuracy

(response correctness) and comprehension time, where reduced accuracy and extended durations suggest elevated cognitive load (Chen et al. 2016; Abbad-Andaloussi et al. 2023; Schreiber et al. 2023).

Physiological measures

Eye-tracking technology allows the continuous monitoring of ocular movements while a user performs a task (Holmqvist et al. 2011). Current eye trackers collect eye-gaze and pupillary measures with high precision, providing high-resolution temporal data, and, thereby, offering precise estimates of cognitive load (Holmqvist et al. 2011; Just and Carpenter 1980; Weber et al. 2021). Fixation detection algorithms can derive fixations (Holmqvist et al. 2011)—periods of time during which the eyes remain stationary on a specific target—from the collected gaze data. The eye-mind hypothesis proposes that fixations reflect ongoing cognitive processing (Just and Carpenter 1980). In (Glöckner and Herbold 2008), the authors identified fixations between 250-500ms as indicating mental processing, while those exceeding 500ms suggest deep mental processing. The eye tracker can also perform a pupillography, i.e. monitoring of the pupil, while eye movements are monitored. Then, pupillary metrics can be used to derive measures such as the low/high index of pupil activity (LHIPA) (Duchowski et al. 2020). This index is measured based on the ratio of low to high frequencies (LF/HF) in pupil size (of the diameter) oscillations, where reduced values correspond to increased cognitive load. Beyond these measures, the LHIPA offers a robust alternative by combining both low- and high-frequency (LF/HF) components of pupil-size oscillations, thereby avoiding reliance on the phasic (LF) response alone (Duchowski et al. 2020). With its formulation based on this LF/HF ratio, the index remains usable for assessing cognitive load even when illumination varies. Moreover, as expected for a frequency-ratio measure, using the LHIPA reduces the need for additional preprocessing of baseline pupil data (Peysakhovich et al. 2015; Duchowski et al. 2020).

Behavioral measures

Eye-tracking data enables investigation of visual behavior from eye movement patterns (Holmqvist et al. 2011; Chen et al. 2016; Weber et al. 2021). Areas of Interest (AOIs) represent a fundamental concept in the analysis of behaviors, denoting stimulus regions containing relevant information (Holmqvist et al. 2011). Each process model element constitutes a distinct AOI in our study. The AOI run count, which consists of tallying AOI entries and exits, was used in (Bera et al. 2019), to measure visual associations between AOIs. This measure estimates the mental effort required to integrate information across several parts of the model, hence, serving as a cognitive load indicator. A similar measure designed for Empirical Software Engineering has shown values well aligned with users' cognitive load (Hao et al. 2023).

While physiological and behavioral measures are particularly well-suited for continuous, fine-grained monitoring of cognitive load, they also complement the broader insights offered by performance and subjective measures.

Building on these complementary strengths, prior research has explored cognitive load at varying levels of granularity. In the process modeling literature, most studies have conducted coarse-grained analyses of cognitive load (Abbad-Andaloussi et al. 2023; Schreiber et al. 2023, 2024; Wang et al. 2022; Wang 2019; Abbad-Andaloussi et

al. 2024; Franceschetti et al. 2024; Figl et al. 2015; Figl et al. 2024; Figl and Laue 2011; Winter et al. 2022; Dobesova and Malcik 2015), typically testing the effect of factors such as process model size (Abbad-Andaloussi et al. 2023), density (Abbad-Andaloussi et al. 2023), separability (Abbad-Andaloussi et al. 2023), model structure (Figl and Laue 2011; Winter et al. 2022), variability of model elements (Figl and Laue 2015; Abbad-Andaloussi et al. 2023), layout (Figl et al. 2024) linked rule annotations (Wang 2019; Wang et al. 2022), modeling knowledge (Figl and Laue 2015) task type (Dobesova and Malcik 2015; Schreiber et al. 2023, 2024; Abbad-Andaloussi et al. 2024), and ambiguity (Abbad-Andaloussi et al. 2025) on users cognitive load at the level of the entire process model or comprehension task. Only a few exceptions have conducted fine-grained analyses based on physiological and behavior measures, focusing on temporal aspects associated with process model comprehension (Schreiber et al. 2024) and sensemaking phases (Chen et al. 2025). However, these studies did not focus on the complexity of process models, which we take as a core subject in this paper.

Visual behavior analysis

Visual behaviors associated with spatial AOIs

Eye tracking allows monitoring changes over time that are related to user's cognitive load and behavior (Holmqvist et al. 2011). Visualizations such as plots of AOIs over time (Holmqvist et al. 2011) allow us to show the sequence of AOIs visited by a user performing a task. When applied in process comprehension, process model elements can be defined as AOIs. Therefore, the visualization displays the sequence of process model elements visited by the user while performing the task. This sequence reflects the multiple visits and revisits to AOIs, yet offering insights into user behavior (Holmqvist et al. 2011). We define two key measures to support this analysis. Namely, the *AOI visits*, which refer to the duration between entering and exiting an AOI (Holmqvist et al. 2011), and the *AOI revisits*, which denote the frequency of returning to an AOI during task execution (Holmqvist et al. 2011).

We derive the *revisit ratio* from these measures: $\rho_{revisit} = \frac{\#revisits}{\#total\ visits}$. The ratio takes values in the range from 0, for pure discovery (no revisits), to 1, when the number of revisits is large. In this study, we use both the AOIs-over-time plot and the revisit ratio to investigate users' behavior (cf. Section [Study Design](#)).

Visual behavior associated with temporal AOIs

Building on the notion of temporal segmentation proposed by (Gidlöf et al. 2013), we further complement the spatial analysis of gaze data by introducing *temporal AOIs*. In their study on decision making in real-world contexts, the authors posited that task progression can be divided into distinct temporal phases derived from eye-tracking information. Specifically, they defined the *first access* to the item relevant for the task—i.e., the moment at which gaze data first indicates that the user has fixated on the relevant element—as a temporal boundary. Using this event as a cutting point, they then identified successive cognitive phases of *orientation* (search), *evaluation* (decision), and *verification*. Adopting this principle, we define temporal AOIs to capture shifts in visual behavior over the course of a comprehension task. In our context, the verification phase described by (Gidlöf et al. 2013) was not explicitly observable, as the task itself required participants to perform a verification; therefore, our temporal AOIs primarily

correspond to the orienting (search) and evaluating (decision) phases. This temporal framing allows the examination of how users' visual attention evolves across the different stages of comprehension, complementing the spatial perspective traditionally applied in BPM eye-tracking studies.

Visualizing behavior in cognitive processes

The evolution of users' visual behavior when comprehending process models has been investigated in several studies (Winter et al. 2023; Djurica et al. 2025; Abbad-Andaloussi et al. 2020; Fındık-Coşkunçay and Ćakır 2022). These studies have, for instance, explored how spatial visual behavior can be represented by projecting eye-tracking data onto process artifacts like process models and decision tables. In (Winter et al. 2023), the authors specifically employed heatmaps and scan-path visualizations to uncover the recurring visual patterns users exhibit during model comprehension. Similarly, (Djurica et al. 2025) examined scan-path visualizations to understand how users search through decision tables summarizing control-flow scenarios. Other studies highlighted temporal visual behavior, emphasizing the order and duration of users' fixations over different model elements. For example, (Abbad-Andaloussi et al. 2020) and (Fındık-Coşkunçay and Ćakır 2022) used scarf plots to depict how readers sequentially engaged with different model elements, how their attention shifted and how they revisited specific model parts over time. These studies elicited different reading strategies and provided valuable insights into process model comprehension.

Visual behavior related to comprehension of BPM process

Within the field of Business Process Management (BPM), prior research has investigated the cognitive mechanisms involved in process model comprehension using eye-tracking techniques. In (Bera et al. 2019), the authors examined how cognitive integration and, more generally, cognitive load associated with task performance (Sweller et al. 1998) can be inferred from gaze-based indicators such as fixation count and AOI run count. More recently, (Winter et al. 2023) investigated gaze patterns in process model reading tasks conducted under controlled experimental conditions. The authors identified *visual routines*—recurring gaze patterns—as manifestations of distinct cognitive activities or processes; their occurrence over time may, in turn, suggest a segmentation of the task into phases. Both studies provide coarse-grained insights into process model comprehension, focusing primarily on results at a task-level. However, they do not fully exploit the potential of eye-tracking to capture fine-grained, high-resolution data (in space and time) that can be directly mapped onto process model structures. Yet these studies leave open questions regarding how gaze data can be examined at multiple levels of granularity. In particular, they suggest the need for approaches that can relate visual behavior to both essential and accidental complexity dimensions of process models.

Existing research provides valuable insights into users' visual behavior during process model comprehension but remains largely confined to task-level analyses. What is still missing is an integrated perspective that considers how gaze behavior evolves both spatially—across different regions of a model—and temporally—throughout the course of a comprehension task. Combining these two dimensions offers the conceptual basis for investigating how process model complexity interacts with cognitive load, thereby extending the methodological foundation for future studies on visual behavior in BPM.

Complexity metrics and cognitive load

This section presents the set of metrics employed to quantify both essential and accidental complexities in the analyzed process models. We selected these metrics for their theoretical relevance to Cognitive Load Theory (Abbad-Andaloussi 2023) (CLT, cf. Section [Cognitive Load Theory, Measurements and Granularity](#)) and their established use in prior research on process model complexity (Mendling 2007; Mendling et al. 2012; Burattin et al. 2017). While Section [Complexity of Process Models and Comprehension Tasks](#) introduced the conceptual categories of essential and accidental complexities, the present section details the specific metrics retained for the study (cf. Section [Study Design](#)) and discusses how each of them relates to cognitive load. Additional details appear in our online supplement³.

Essential complexity metrics

This first category of metrics encompasses structural characteristics of process models that inherently affect the cognitive load required for comprehension. These metrics capture properties that contribute to the intrinsic cognitive load associated with understanding the model's logic and behavior (Mendling 2007; Mendling et al. 2012).

Size

The number of nodes in the model (e.g., tasks, gateways, events) (Mendling 2007; Mendling et al. 2012). The number of nodes contained in a process model can affect how easily the model is understood. According to cognitive load theory, human working memory has a limited capacity for processing information, which constrains the amount of detail individuals can effectively manage at once. Consequently, a model with a large size (i.e. a large amount of information) can increase the intrinsic cognitive load experienced during comprehension (Sweller et al. 2011; Chen et al. 2016).

Density

Relates the number of edges (possible flows) to the size of the model (Mendling 2007; Mendling et al. 2012). Multiple possible flows within a process model can challenge its interpretation, as it requires additional cognitive load to trace and verify relationships between elements. From a cognitive load perspective, increasing the density of a model results in an expansion of the amount of information that readers must process simultaneously, thereby increasing the intrinsic cognitive load experienced during model comprehension.

Partitionability

Considers the relationship of sub-components to the overall model (Mendling 2007; Mendling et al. 2012). Partitionability refers to the user's ability to isolate or group a specific sub-part of a model. It facilitates the comprehension of the model by focusing on this sub-part as a single, meaningful unit rather than as many individual elements. This ability creates a higher level of abstraction where users only need to think about how this sub-part connects to the rest of the model, instead of how each individual element relates to the model. This, in turn, reduces their intrinsic load.

³ Online appendix: <https://doi.org/10.5281/zenodo.17991805>

Connector interplay

Considers the interactions and effects of the different connector types (Mendling 2007; Mendling et al. 2012). In addition to their number, the type of connectors (which may be referred to as *gateways*) used also affects the complexity of a model. In the present study, we focus on models that do not present errors. Therefore, we do not address the concept of connector mismatch in this context. The plurality of connectors—reported by the Control Flow Complexity (CFC) and their Connector Heterogeneity (CH)—requires a greater cognitive effort to evaluate the different possible states of a model, resulting in an increase in intrinsic cognitive load.

Cyclicity

Captures the number of nodes for which a cycle exists, then provides the ratio of this number to the total number of nodes of the model (Mendling 2007; Mendling et al. 2012). Cycles have the potential to increase the size of a model, which, in turn, introduces a complexity that impacts the intrinsic load due to the limitations of users' working memory.

Concurrency

Explores the possible concurrent paths of a model. The Token split metric counts the control tokens (e.g. AND or OR) incorporated in the model (Mendling 2007; Mendling et al. 2012). The use of connectors or gateways such as AND or OR splits a path into new control threads, i.e., concurrent paths. In order to understand a model, the user must be able to keep track of the various threads. Therefore, because of the limited human working memory, understanding a model comprising a large number of threads can also increase the intrinsic cognitive load.

Accidental complexity metrics

This second category comprises layout- and representation-related characteristics that influence the extraneous cognitive load experienced during model comprehension. These metrics reflect visual and graphical factors that may facilitate or hinder the user's ability to interpret the model efficiently (Bernstein and Soffer 2015; Burattin et al. 2017).

Edges style

A measure of the style of the edges as the ratio of simple (default) or 'broken' (with breaking points) edges to the total number of edges (Bernstein and Soffer 2015). The graphical aspects of edges captured by this category of metrics contributes to make the identification of patterns in the model more difficult. Increasing the length of the connectors and the number of segments that compose them results in distortion of the representation of the patterns. In order to correctly identify patterns, users must invest mental effort to compensate for such distortion, thereby increasing the extraneous load.

Crossing edges

This metric denotes the ratio of the number of crossing edges to the total number of edges (Bernstein and Soffer 2015). A model designed with multiple crossed edges may require the user to disentangle the edges in order to correctly follow the process flow or

identify patterns within the model. Such mental rearrangement impacts the extraneous load.

Angles

This metric denotes the ratio of the number of *orthogonal segments* to the total number of segments (Bernstein and Soffer 2015), where *orthogonal segments* are parts of edges which are aligned with a grid layout of the model. In a similar manner, when the angles of edges tend to be arbitrary, the user must proceed to a rearrangement of the mental representation, which also impacts extraneous load.

Symmetry in blocks

This metric denotes the symmetry of the elements' arrangement inside a block of the model (Bernstein and Soffer 2015). In (Bernstein and Soffer 2015), the authors derived the notion of a block from a qualitative study and does not provide an objective definition that would allow automated detection or measurement. Since our analysis considers models at different levels of granularity, such a concept remains relevant to our investigation, but requires a more formalized interpretation. Therefore, we adopt a refined definition that can be used in an automated detection process, whereby a block refers to all nodes (connectors, activities, edges, etc.) that we place between matching connectors. As noted by (Abbad-Andalousi 2021), symmetry, as a modeling pattern, can facilitate the representation of blocks within a process model. The presence of symmetry in model blocks can thus help users to construct abstract mental representations and to access information more efficiently. This phenomenon corresponds to a reduction in extraneous cognitive load, and, hence, in the required cognitive effort for the comprehension.

Flow consistency

Metrics in this category capture the extent to which a process model maintains a coherent directional flow from its start to its end (Burattin et al. 2017). In (Burattin et al. 2017), the authors investigated several variants of flow consistency metrics, among which we selected the activity positioning-based variant for this study, as it demonstrated the best performance. In addition, it completes our set of metrics by introducing a new dimension related to activities. When the measure of flow consistency is high, models preserve a clear and stable directionality, which facilitates users' ability to follow the process progression and thereby reduces the extraneous cognitive load required to construct an accurate mental representation of the model.

We selected the metrics presented above to capture the main sources of complexity relevant to process model comprehension. In the following section, they serve to operationalize the theoretical constructs of essential and accidental complexity defined in our study design.

Research method

This section describes the methodology used for our investigation (cf. Section [Study Design](#)). Then, it outlines the experimental procedure (cf. Section [Experimental Procedure](#)) and, finally, presents the methodologies for data collection and data analysis (cf. Section [Data Collection and Analysis Procedure](#)).

Study design

Section [Research Model](#) introduces our research model. Then, Section 4.1.2 details the material used during the eye-tracking experiment, while Section [Participants](#) presents the participants.

Research model

To address the overarching research objective of this work (i.e., understanding how process model complexity influences users' cognitive load), we designed a two-tiered analytical strategy:

- **Coarse-grained analysis:** Demonstrates the impact of essential and accidental complexity on cognitive load at the level of entire process models and complete tasks.
- **Fine-grained analysis:** Investigates users' visual behavior and cognitive load over time and space while interacting with process models.

This division enables the integration of high-level insights from the coarse-grained analysis with detailed observations of behavioral patterns and cognitive processing from the fine-grained analysis. Together, these complementary perspectives offer a more comprehensive understanding of the relationship between process model complexity and users' cognitive load.

Coarse-grained analysis The coarse-grained analysis focuses on assessing how the complexity of entire process models affects users' cognitive load. We divided the analysis into two complementary hypothesis-testing sub-models:

Model-level complexity Investigating the impact of essential and accidental complexity in full models.

Task-level complexity Examining how targeted tasks involving specific model parts affect this relationship.

This analysis allows us to address two related research questions at a higher abstraction level, before zooming into the *fine-grained analysis*.

Hypothesis testing on model complexity The first sub-model investigates the relationship between overall process model complexity and users' cognitive load. It assesses whether models characterized by higher essential or accidental complexity result in greater cognitive effort during comprehension.

Figure 1 illustrates the research model. In the *Treatment*, we define essential and accidental complexity as the theoretical constructs. We operationalized these constructs through process models that we classified as simple or complex, using the metrics in Sect. [Complexity of Process Models and Comprehension Tasks](#). The *Output* of the model is cognitive load, operationalized via the measures introduced in Sect. [Cognitive Load Theory, Measurements and Granularity](#). We test the following hypotheses:

- **H1** Models with high essential complexity yield higher cognitive load than models with low essential complexity.
- **H2** Models with high accidental complexity yield higher cognitive load than models with low accidental complexity.

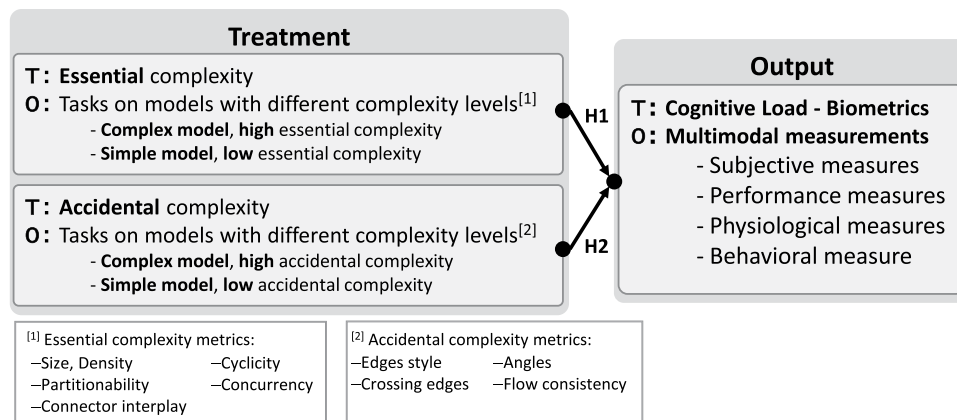


Fig. 1 Research model for hypothesis testing at the model level. **T:** Theoretical construct. **O:** Operationalization. Full-resolution version available in the online appendix (<https://doi.org/10.5281/zenodo.17991805>)

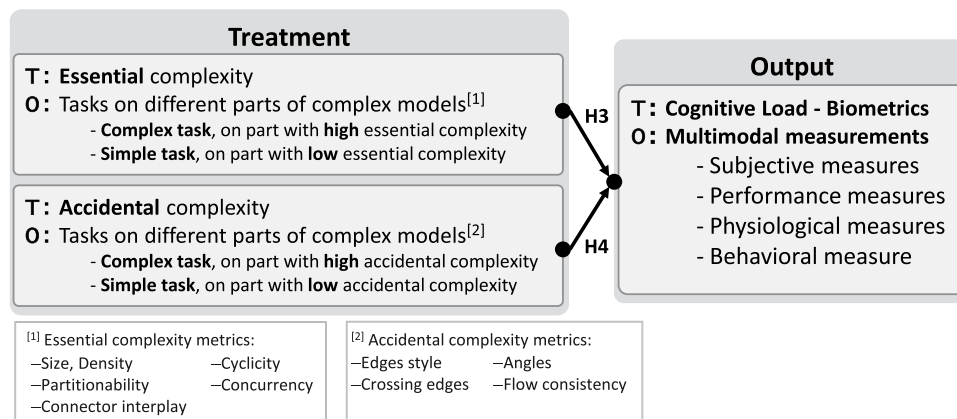


Fig. 2 Research model for hypothesis testing at the task level. **T:** Theoretical construct. **O:** Operationalization. Full-resolution version available in the online appendix (<https://doi.org/10.5281/zenodo.17991805>)

To investigate these hypotheses, participants complete comprehension tasks based on full models exhibiting either low or high complexity. The corresponding cognitive load measures are then compared across these two levels. This experimental design enables us to address **RQ1**: *Does process model complexity as captured by existing complexity metrics align with measures of users' cognitive load?*

Hypothesis testing on task complexity The second sub-model refines the perspective by introducing task complexity as a factor that might affect the relationship between process model complexity and users' cognitive load. This sub-model examines specific tasks that target subparts of models with varying complexity.

The goal is to investigate whether the interaction with more or less complex regions of a model (via tasks on subparts of the model with varying complexity) affects users' cognitive load, even when the overall model remains the same.

Figure 2 presents the research model. As in the previous sub-model, the theoretical constructs in the *Treatment* are essential and accidental complexity. However, here we operationalized them through tasks targeting model parts of low or high complexity. The *Output*, once again, is cognitive load, measured using the same measures described earlier. We formulate the following hypotheses:

- **H3** Complex tasks addressing parts of the model with high essential complexity yield higher cognitive load than simple tasks addressing other parts of the same model with low essential complexity.
- **H4** Complex tasks addressing parts of the model with high accidental complexity yield higher cognitive load than simple tasks addressing other parts of the same model with low accidental complexity.

In the experiment, we asked participants to complete comprehension tasks focusing on specific model parts differing in complexity. The resulting cognitive load data is compared across the two task complexity levels. The results from this analysis will contribute to answering **RQ2**: *Does the complexity of tasks affect the alignment between existing complexity metrics and measures of users’ cognitive load?*

Fine-grained analysis The fine-grained analysis adopts an exploratory approach based on the Goal-Question-Metric (GQM) paradigm (Basili et al. 1994). It investigates users’ comprehension behavior at a fine-grained level by analyzing their visual attention patterns over time and across spatial AOIs within process models. This part of the analysis addresses three research questions (**RQ3–RQ5**) across two sub-models (summarized in Table 1).

Analysis in the time perspective only The first analysis focuses on the temporal evolution of users’ visual behavior as they comprehend process models of varying complexity. It aims to uncover how comprehension unfolds over time. Thus, answering **RQ3**: *How do users’ visual behaviors evolve over time during comprehension tasks on process models of varying complexity?*

This model explores eye-tracking data by analyzing AOI visit timelines and fixation durations. It also includes computed measures such as comprehension time, number of deep mental processing fixations, and revisit ratio. These measures provide insight into the temporal dynamics of cognitive processing during model comprehension.

Analysis in space and time perspectives This second exploratory analysis combines both spatial and temporal aspects of users’ interaction with process models. It examines *whether* and *how* complexity in non-task-relevant areas of a model (i.e., those not directly

Table 1 Summary of research goals, questions, and associated measures used in the fine-grained analysis, structured according to the Goal-Question-Metric (GQM) template (Basili et al. 1994)

Goal	Purpose	Explore users’ visual behavior at a fine-grained level
Question	Issue	during comprehension tasks
	Object	using simple and complex process models
	RQ3	<i>How do users’ visual behaviors evolve over time during comprehension tasks on process models of varying complexity?</i>
Measures		Comprehension time, number of deep mental processing fixations, revisit ratio
Question	Purpose	Explore users’ cognitive load at a fine-grained level
	Issue	during comprehension tasks
	Object	using spatial and temporal segmentation of eye-tracking data
Measures	RQ4/RQ5	<i>Does the complexity of the non-task-relevant parts of process models affect measures of users’ cognitive load during comprehension tasks?</i> <i>How does the complexity of the non-task-relevant parts of process models affect users’ cognitive load during different phases of comprehension tasks?</i>
		AOI run count, number of mental and deep mental processing fixations

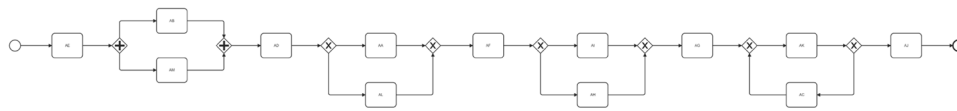


Fig. 3 Models of set A – Simple

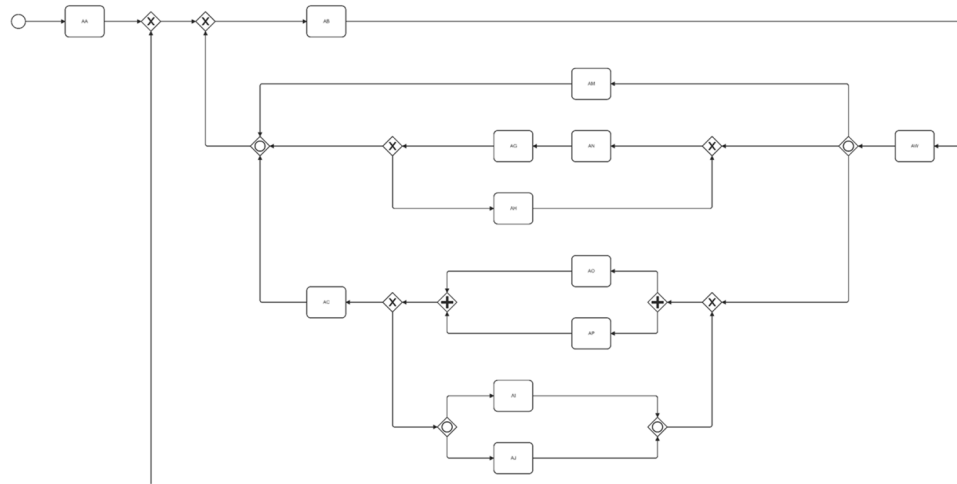


Fig. 4 Models of set A – Complex (with essential complexity). Due to space limitations, we show this model partially

involved in the task) affects users' cognitive load. The associated research questions are: **RQ4:** *Does the complexity of the non-task-relevant parts of process models affect measures of users' cognitive load during comprehension tasks?* and **RQ5:** *How does the complexity of the non-task-relevant parts of process models affect measures of users' cognitive load during comprehension tasks?*

To address these questions, we analyze the distribution of AOI run count, mental processing and deep mental processing fixations on task-relevant and non-task-relevant parts (for **RQ4**) as well as in discovery and post-discovery phases (for **RQ5**).

Material

To investigate our hypotheses and address the research questions, we designed two sets of process models: Set A and Set B. Each set serves specific analytical purposes: Set A supports analyses related to **RQ1** and **RQ3**; Set B supports analyses related to **RQ2**, **RQ4**, and **RQ5**.

Set A This set consists of three distinct models: a simple model (Fig. 3) with low essential and accidental complexity, and two complex models: one with high essential complexity (Fig. 4) and one with high accidental complexity (Fig. 5). The simple model and the high essential complexity model differ in structural characteristics including size, density, partitionability, connector interplay, cyclicity, and concurrency (cf. Fig. 1). In contrast, the simple model and the high accidental complexity model differ in layout-related aspects, including edge style, crossings, angles, block symmetry, and flow consistency.

Set B This second set comprises two composed models: one representing essential complexity (Fig. 6) and one representing accidental complexity (Fig. 7). Each composed model

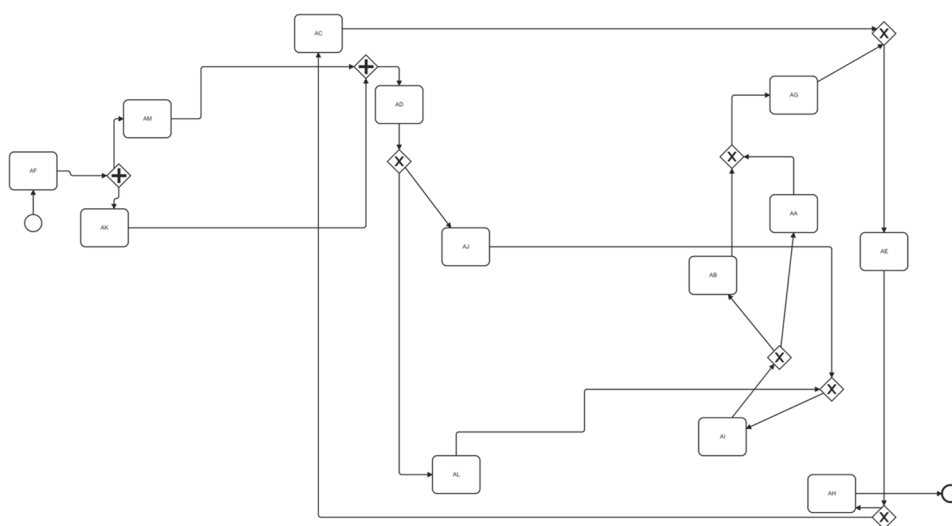


Fig. 5 Models of set A – Complex (with accidental complexity)

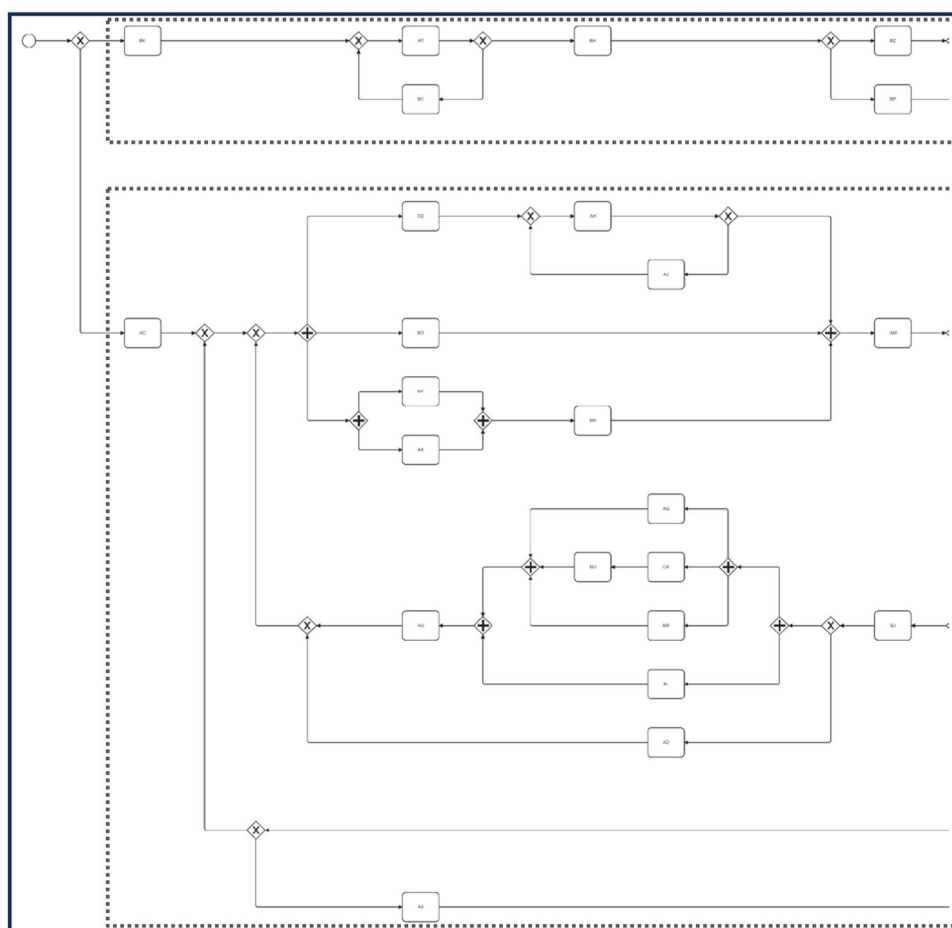


Fig. 6 Models of set B. Model with essential complexity – simple part, on top, and complex part at the bottom.

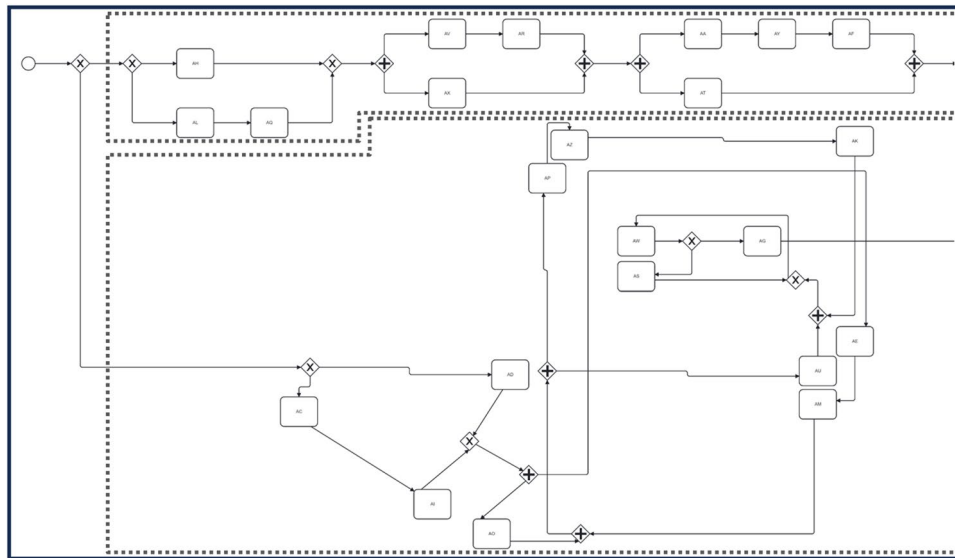


Fig. 7 Models of set B. Model with accidental complexity – simple part, on top, and complex part at the bottom.

is structured into two parts: a simple part (low complexity) in the upper part of the model, and a complex part (high complexity) in the lower part. For the model with essential complexity, the complex part differs from the simple part in terms of structure (i.e., size, density, partitionability, connector interplay, cyclicity, and concurrency). For the model with accidental complexity, the difference lies in visual layout (i.e., edge style, crossings, angles, block symmetry, and flow consistency). These distinctions enable localized task design targeting either simple or complex regions within the same model (cf. Fig. 2).

Activity labeling To ensure that participants' performance reflected cognitive processing of the model structure rather than prior familiarity with a specific domain, we deliberately controlled the labeling of activities. In (Yan et al. 2017), the authors proposed a technique for generating large collections of models. This approach involved the creation of synthetic models or tasks that simulated realistic process structures while remaining domain-neutral. Nevertheless, a potential drawback of such automatically generated models was the possibility of introducing artificial or confusing labels for users. Previous research had shown that the style and content of activity labels could substantially influence comprehension. For instance, (Mendling et al. 2010) demonstrated that using verb–noun labels led to better model understanding compared to alternative labeling conventions, highlighting how semantic content shaped interpretation. However, (Wohlin et al. 2024) emphasized that domain knowledge could act as a confounding factor, potentially biasing the assessment of cognitive effort. Similarly, (Schreiber et al. 2023) employed generic, layman's terminology in their models to avoid the impact of process-specific vocabulary, while (Bernstein and Soffer 2015) went further by blurring all textual elements to direct attention exclusively toward the visual layout. Since this study investigated the impact of process model complexity on cognitive load, we sought to avoid situations where the complexity of textual elements could interfere with the complexity of graphical elements. In the context of this study, to minimize the influence of domain knowledge, all activity labels were anonymized using random two-letter codes.

All process models use core BPMN 2.0 elements (start/end events, activities, gateways, and edges) and incorporate common workflow patterns such as sequence, concurrency, exclusive/inclusive choice, and repetition (Muehlen and Recker 2008). This element set ensures realistic model behavior while enabling controlled variation in complexity. We achieved high essential and accidental complexities by systematically modifying base models or subparts until all relevant metrics indicated a meaningful increase. The final process models used in the experiment are available in our online appendix⁴.

Task design As depicted in Fig. 8, each comprehension task consists of a unique question (“Given this model, is the following assertion valid?”) to validate an assertion about a given model. To complete the task, the participant is required to answer the question by selecting *yes*, *no*, or *I don’t know*. The list of all tasks used in the experiment is available in our online appendix⁵.

Assertions Table 2 presents the different types of assertions used for models of Set A (upper portion of the table) and Set B (lower portion of the table). For each type (first column), an example is given (second column) as well as the designation of the principal control flow pattern that captures this type of assertion.

We designed four distinct types of assertions related to different control flow patterns (Order, Choice, Repetition, Concurrency). Each type can be used to generate a new assertion by replacing the labels of the activities involved. To avoid confounding factors, assertions feature exactly two activities, which we have positioned at a fixed graphical distance from each other, and which we have formulated in positive terms (to avoid additional cognitive load resulting from negation). For example, we obtain the following sentence from a “Choice” assertion involving activities with labels AL and AC: “In this process, if activity AL is executed, activity AC never occurs, and conversely.”

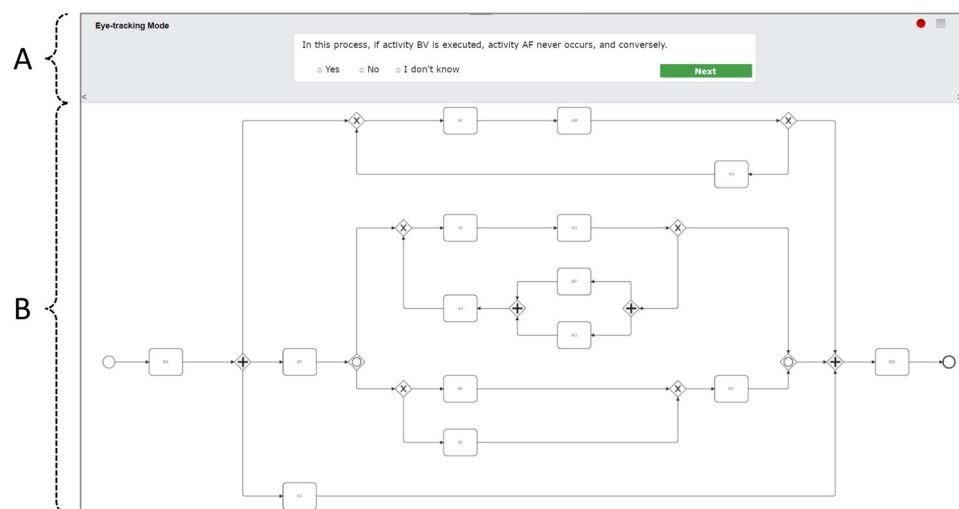


Fig. 8 Example of a complete task. In part A, we present the question as an assertion that participants must validate against the provided answers. In part B, the model is displayed in a navigable area (with movements in four directions and zoom in and out)

⁴ Online appendix: <https://doi.org/10.5281/zenodo.17991805>

⁵ Online appendix: <https://doi.org/10.5281/zenodo.17991805>

Table 2 List of assertions. An assertion always involves two activities. The label of an activity is a random concatenation of two letters

Type	Example	Control flow
Assertions for models of Set A		
Choice	In this process, if activity BA is executed, activity AZ never occurs, and conversely.	Exclusive gateways
Repetition	This process has an execution where activity BC occurs several times while activity AK occurs only once.	Exclusive gateways, used to create a loop-back sequence
Assertions for models of Set B		
Order	In every execution of this process, there is an occurrence of activity BL before any occurrence of activity AT.	Sequence
Choice	In this process, if activity AP is executed, activity BN never occurs, and conversely.	Exclusive gateways
Concurrency	In this process, activity BU is executed in parallel with activity AS.	Parallel gateways

Assertions for models in Set A allow a total of six unique tasks to be created. Using the same types of assertions across all models allows to ensure the comparability across conditions (varying complexity types and levels) and limit the introduction of potential bias.

Regarding the assertions for models in Set B, we associated two out of three types with each model. Since the models are composed, each assertion has two variants—one for the simple part and one for the complex part—for a total of eight unique tasks. As mentioned in Sect. [Complexity of Process Models and Comprehension Tasks](#), we refer to tasks targeting the simple part of a composed model as *simple tasks*, and tasks targeting the complex part as *complex tasks*.

Participants

Our experiment involved twenty-seven individuals with academic backgrounds in Computer Science or Information Systems, recruited from the University of St. Gallen. The participant pool included fourteen subjects (11 male, 3 female) reported prior knowledge of BPMN, while the remaining thirteen (8 male, 5 female) possessed no prior knowledge of this modeling standard. Participants' ages averaged approximately 30 years ($\mu = 31.19, \sigma = 7.31$).

Experimental procedure

The within-subject experiment followed a standardized protocol in controlled laboratory conditions. Sessions began with a comprehensive familiarization phase covering: (1) the relevant BPMN subset employed to design the material (cf. Section [Task-Relevant Versus Non-Task-Relevant Analysis](#)), (2) all assertion types used in tasks, and (3) practical exercises with support and feedback. The exercises took the form of four tasks similar to the experiment, with participants receiving assistance and guidance throughout. After the exercises, we initiated a discussion in which we examined each response to ensure comprehension. Subsequently, the participants were requested to evaluate the model understanding difficulty (self-assessment) of each task using a 5-point Likert scale. Participants then completed a demographic survey and calibration procedure.

To ensure data quality, we implemented multiple safeguards against common human factors resulting in eye-tracking artifacts (Holmqvist et al. 2011). These included controls of a correct posture with the eye-tracker, randomization of task sequences for each participant to prevent potential learning effects, and an initial practice task (excluded

from analysis, though the participants are not told it is a dummy task) for a further familiarization with the recording equipment. Throughout the session, participants repeated the following two steps for each task: first, they completed the task (i.e., to validate an assertion on a provided model), and then they evaluated the difficulty of understanding the model via a 5-point Likert scale.

The systematic randomization of tasks and the provision of detailed information typically contributed to limit the participant's germane load. These measures limit the processing of new information to what is required for task execution, thereby reducing the necessity for new mental schemes and, hence, maintaining the germane load stable.

Data collection and analysis procedure

Data collection

Data collection proceeded through a preliminary phase of surveys (unrecorded) followed by the experimental trials (recorded). The core recording equipment consists of an eye-tracker, Tobii X3-120, mounted on a 24" monitor (1920x1080 resolution) under controlled light conditions to mitigate recording issues (Holmqvist et al. 2011), and managed through the EyeMind platform (Abbad-Andaloussi et al. 2023).

Light conditions during data collection Because pupil size is highly sensitive to lighting conditions and key to eye-tracking accuracy (Holmqvist et al. 2011), we controlled the visual environment to ensure stable measurements throughout the experiment. All recordings were conducted in the same laboratory room using identical hardware (a 24" monitor with fixed luminance settings) and under consistent artificial lighting. In detail, we collected data in a room with windows obscured by blinds. Furthermore, we adjusted the ambient light in the room to 200 lux at the desktop where the participants were seated, as well as the luminance of the screen in order to obtain an illumination in the range of 50 to 80 lux at the participant's face. These measures helped maintain uniform light levels both within each participant's session and across participants. We further ensured that navigation within the BPMN models (scrolling, zooming, sub-process expansion) did not introduce meaningful luminance variation. The diagrams consist predominantly of a white background with thin black contour lines, and we observed that even at extreme zoom levels the proportion of darker pixels remains negligible. We computed the luminance of the navigable part of the interface (part B in Fig. 8) for all models at normal zoom levels (i.e., when the models completely appear within the interface), as well as at extreme zoom levels at parts with a high density of BPMN elements. With our settings, the total variation in the luminance of the interface did not exceed 4%, with a mean at 98.63% and a standard deviation of 1.44% (the screen captures and the mean luminance graph are available in our online appendix⁶). As a result, navigation changes only the geometric framing of the same visual content without affecting the overall on-screen luminance. The interface was displayed in full-screen mode, ensuring that no other on-screen content could alter the luminance perceived by the eye tracker.

Data collection using EyeMind We relied on EyeMind—an open-source platform specifically designed for eye-tracking studies on large and interactive process models—to

⁶ Online appendix: <https://doi.org/10.5281/zenodo.17991805>

control the experiment and to automate the mapping of gaze data to process model elements. While static stimuli (e.g., non-interactive images of small models) facilitate analysis by keeping AOI locations fixed on the screen (Holmqvist et al. 2011), dynamic stimuli were essential for our setting to accommodate realistic model scale and complexity, support natural navigation behaviors (scrolling, zooming, and sub-process exploration), and thereby enhance ecological validity. EyeMind addresses the technical challenges of dynamic presentation by capturing a screen snapshot after every user interaction and recalculating the AOI coordinates accordingly, ensuring accurate gaze-to-AOI mapping even when the model is moved or rescaled (Abbad-Andaloussi et al. 2023).

Data analysis

We divided the analysis of the collected data into two main parts, reflecting the dual structure of our study design (cf. Section [Research Model](#)). Each part corresponds to a different level of analysis (i.e., coarse-grained or fine-grained) and is associated with a specific subset of research questions. The following paragraphs summarize the applied procedures; comprehensive details of the analysis are available in the Python notebooks included in our online appendix⁷.

Coarse-grained analysis This part of the analysis addresses **RQ1** and **RQ2**, which investigate how model-level and task-level complexity affect cognitive load.

For **RQ1**, we analyzed data from **Set A**, excluding recordings affected by eye-tracking issues. From a theoretical maximum of 200 recordings (2 questions $\times$ 3 models $\times$ 25 participants), we retained 149 usable recordings across 25 participants. We excluded two participants entirely due to calibration problems, and one participant missed a single task due to an erroneous manipulation in the user interface.

We calculated cognitive load measures as defined in Sect. [Cognitive Load Theory, Measurements and Granularity](#) and computed both descriptive and inferential statistics. Specifically, we used the non-parametric Wilcoxon Signed-Rank paired test (single-tailed) to assess whether mean differences in cognitive load indicators between low- and high-complexity models were statistically significant. This test is robust against non-normal data distributions.

A similar process was applied to address **RQ2**, using data from **Set B**. With 25 participants completing all tasks, we obtained 200 recordings (2 questions $\times$ 4 tasks $\times$ 25 participants). Cognitive load measures and statistical testing followed the same procedure as for **RQ1**.

Fine-grained analysis This part of the analysis investigates participants' visual behavior and cognitive load during comprehension tasks. It adopts a descriptive and exploratory approach, complemented by inferential statistical analysis to support the findings. It addresses **RQ3**, **RQ4**, and **RQ5**.

We analyzed **RQ3** using data from **Set A**. We generated three types of visualizations for each task instance (trial): (1) sequences of AOIs visited over time (AOI timeline plots) (Holmqvist et al. 2011), (2) AOI timelines with overlaid patterns and trend highlights, and (3) plots showing the evolution of fixation durations over time.

⁷ Online appendix: <https://doi.org/10.5281/zenodo.17991805>

Trials were grouped into categories based on model complexity: tasks on the simple model, tasks on the model with high essential complexity, and tasks on the model with high accidental complexity. This categorization allowed us to identify differences in visual behavior across complexity levels (see Fig. 9 for examples).

We analyzed **RQ4** and **RQ5** using data from **Set B**, which includes composed models with both simple and complex parts. We first defined additional spatial AOIs corresponding to these two parts of each model. This enabled the segmentation of fixation data based on whether a fixation occurred on the simple or complex part.

We collected 200 recordings from 25 participants, each completing 4 tasks on 2 composed models⁸. We then computed cognitive load indicators (cf. Sect. [Cognitive Load Theory, Measurements and Granularity](#)) and conducted a Wilcoxon Signed-Rank paired test (single-tailed) to determine whether differences in cognitive load between AOIs (simple vs. complex parts) were statistically significant.

To further support the analysis for each task, we compared cognitive load measures computed separately for fixations on the simple and complex AOIs. We calculated the difference in proportion of indicator values between these AOIs (relative to the total per task). This proportion difference was then compared across task types (i.e., tasks targeting the simple part vs. the complex part of the model).

This analysis allows us to investigate how process model and associated task complexities influence visual behavior and cognitive load at a fine-grained level.

Findings

In this section, we present the findings from the two data analyses separately. Section “[Coarse-grained analysis](#)” details the coarse-grained analysis, while Sect. [Fine-grained Analysis](#) details the fine-grained analysis.

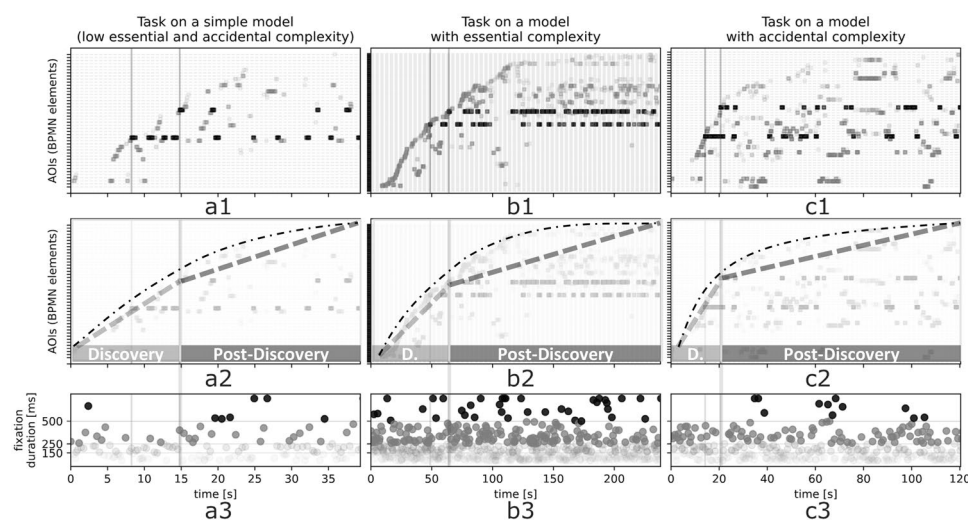


Fig. 9 Visual behavior of participant P22 on three tasks: **(a)** a task on a simple model, **(b)** a task on a model with high essential complexity, and **(c)** a task on a model with high accidental complexity. Top row: AOI visits over time; vertical ticks mark visit order; vertical lines show when task-relevant AOIs were first visited. Middle row: same plots overlaid with envelope and slope estimation lines (gray dashed). Bottom row: fixations over time by duration. This figure is available in the online appendix (<https://doi.org/10.5281/zenodo.17991805>)

⁸200 tasks = 4 questions × 2 models × 25 participants.

Coarse-grained analysis

The subsequent sections present the results for the coarse-grained analysis. Section [Impact of model complexity on cognitive load](#) outlines the results of the analysis on the impact of model complexity on cognitive load. Then, Section [Impact of model complexity on cognitive load](#) presents the results of the analysis on the impact of task complexity on cognitive load.

Impact of model complexity on cognitive load

This part of the analysis addresses **RQ1**: *Does process model complexity, as captured by existing complexity metrics, align with measures of users' cognitive load?* To answer this question, we used data from **Set A**, which includes one simple model and two complex models differing in either essential or accidental complexity (cf. Section [“Material”](#)).

The left side of Table 3 presents the results related to **H1** (essential complexity), and the right side to **H2** (accidental complexity).

For each condition, we report descriptive statistics (mean values for simple and complex models, and effect size) and results from inferential testing.

The descriptive statistics indicate that models exhibiting high complexity (whether essential or accidental) are perceived as more challenging compared to simple models. Participants demonstrated lower accuracy and longer comprehension times while performing tasks on these complex models. They also demonstrated an increase in AOI run count, in the number of mental processing fixations (with duration in the range 250–500ms), and in the number of deep mental processing fixations (with duration greater than 500ms). The final measure, LHIPA, shows lower values for models with high complexity compared to simple models, which also indicates a high cognitive load when performing tasks on complex models (see Section [Cognitive Load Theory, Measurements and Granularity](#)).

These patterns are consistent with the CLT (cognitive load theory, see Section [Cognitive Load Theory, Measurements and Granularity](#)) and suggest that both essential and accidental complexity contribute to increased cognitive load during model comprehension.

Table 3 Descriptive and inferential statistics for the impact of models' complexity. Abbreviations: CL: Cognitive load; Desc. stat.: Descriptive statistics; Inf. stat.: Inferential statistics, #: Number of observations; Comp.: Comprehension; N.: number; P.: Processing; D.: Deep; LHIPA: Low/High Index of Pupil Activity; M_{Simple} ; $M_{Complex}$: calculated means for models with low and high complexity, respectively; ES: effect size using *Cohen's d*. ES can be categorized into small ($= 0.2$), medium ($= 0.5$), or large (≥ 0.8). A p -value $< .05$ means that the results of pairwise comparisons are significant

Impact of models' complexity															
Measure of CL	H1: Essential complexity					H2: Accidental complexity									
	Desc.	Means			ES	Inf.	Desc.	Means			Inf.				
		#	M_{Simple}	$M_{Complex}$				d	p	#		M_{Simple}	$M_{Complex}$	d	p
Perceived difficulty	50	0.620	2.360	> 1	<.001	50	0.620	2.280	> 1	<.001					
Comp. accuracy	50	0.980	0.660	>.8	<.001	50	0.980	0.880	0.37	.0625					
Comp. time	49	44.938	145.137	> 1	<.001	49	44.938	104.020	> 1	<.001					
AOI run count	49	88.163	232.120	> 1	<.001	49	88.143	204.540	> 1	<.001					
N. Mental P. fix.	49	27.857	85.520	> 1	<.001	49	27.857	71.200	> 1	<.001					
N. D. Mental P. fix.	48	4.250	12.320	> 1	<.001	47	4.250	11.876	> 1	<.001					
LHIPA	48	1.320	0.636	> 1	<.001	48	1.320	0.786	> 1	<.001					

An exception to these patterns occurs with Comprehension Accuracy for hypothesis H2 (comparison between simple tasks and complex tasks with accidental complexity). In this particular case, the observed high p -value ($p = .0625$) indicates that the metric alone is not sufficient to clearly capture the impact of the complexity in this specific context. However, the inferential statistics provide further support for the ensuing interpretation: for nearly all measures, the differences between simple and complex models are statistically significant ($p < .001$), thereby substantiating both *H1* and *H2*.

Impact of task complexity on cognitive load

This part of the analysis addresses **RQ2**: *Does the complexity of tasks affect the alignment between existing complexity metrics and measures of users' cognitive load?* To investigate this, we analyze task-level complexity using composed models from **Set B** (cf. Section [Material](#)). These models include both simple and complex parts, allowing task complexity to be manipulated independently of overall model complexity.

Table 4 presents descriptive and inferential statistics for cognitive load measures related to **H3** (essential complexity) and **H4** (accidental complexity). For each hypothesis, results compare performance on tasks targeting simple parts of the model (low complexity) versus tasks targeting complex parts (high complexity).

The data show that, in both essential and accidental complexity conditions, complex tasks are perceived as significantly more difficult, lead to lower comprehension accuracy, and require longer comprehension time. They also result in higher AOI run counts, more mental and deep mental processing fixations, and lower LHIPA values—indicators of increased cognitive load.

Comprehension accuracy for tasks with accidental complexity shows an exception in its measures (similar to observations in Section “[Impact of model complexity on cognitive load](#)”). The difference between task types, with small practical significance ($d = 0.35$), does not reach statistical significance ($p = 0.173$). This exception further demonstrate that the metric is not adequate to clearly capture the impact of the complexity in this specific context.

Table 4 Descriptive and inferential statistics for the impact of tasks' complexity. Abbreviations: CL: Cognitive load; Desc. stat.: Descriptive statistics; Inf. stat.: Inferential statistics, #: Number of observations; Comp.: Comprehension; N.: number; P.: Processing; D.: Deep; LHIPA: Low/High Index of Pupil Activity; M_{Simple} ; $M_{Complex}$: calculated means for models with low and high complexity, respectively; ES: effect size using *Cohen's d*. ES can be categorized into small ($= 0.2$), medium ($= 0.5$), or large (≥ 0.8). A p -value $< .05$ means that the results of pairwise comparisons are significant

Impact of tasks' complexity										
Measure of CL	H3: Essential complexity					H4: Accidental complexity				
	Desc.				Inf.	Desc.				Inf.
		Means	ES				Means	ES		
	#	M_{Simple}	$M_{Complex}$	d	p	#	M_{Simple}	$M_{Complex}$	d	p
Perceived difficulty	50	1.100	2.280	> 1	$< .001$	50	0.900	1.980	> 1	$< .001$
Comp. accuracy	50	0.940	0.720	$> .5$	$< .001$	50	0.920	0.800	0.35	.173
Comp. time	50	89.857	143.679	$> .8$	$< .001$	50	62.347	95.298	$> .5$	$< .001$
AOI run count	50	156.980	242.540	$> .8$	$< .001$	50	122.640	180.020	$> .5$	$< .001$
N. Mental P. fix.	49	53.800	83.560	$> .8$	$< .001$	50	36.700	62.300	$> .8$	$< .001$
N. D. Mental P. fix.	49	7.327	12.300	$> .5$	$< .001$	49	5.080	10.735	$> .8$	$< .001$
LHIPA	50	0.921	0.679	$> .5$	$< .001$	49	1.143	0.849	$> .5$	$< .001$

Notwithstanding this particular case of comprehension accuracy within accidental complexity, these results indicate that task complexity has a noticeable effect on cognitive load, even when the underlying model remains the same.

Specifically, tasks that target complex areas of a model produce significantly higher cognitive load than tasks focused on simpler parts. This pattern supports both *H3* and *H4*, as confirmed by the inferential statistics ($p < 0.05$) for nearly all cognitive load indicators.

Fine-grained analysis

This section reports the findings from the fine-grained analysis, which is structured along three analytical directions, each addressing a specific research question. Specifically, Section “[Comprehension Phases](#)” presents findings related to comprehension phases (RQ3), Section “[Task-Relevant Versus Non-Task-Relevant Analysis](#)” reports findings on task-relevant versus non-task-relevant processing (RQ4), and Section “[Discovery versus post-discovery analysis](#)” presents findings comparing discovery and post-discovery phases (RQ5).

Table 5 lists the nine observations derived from the fine-grained analysis, grouped according to the corresponding analytical direction.

Comprehension phases

The third part of our analysis addresses RQ3: *How do users’ visual behaviors evolve over time during comprehension tasks on process models of varying complexity?* It focuses on tasks from **Set A**.

To illustrate emerging behavioral patterns, we present eye-tracking data from a single representative participant (P22), who completed tasks on simple models, as well as tasks on models with high essential and high accidental complexity (see Fig. 9). A broader view of results across participants is available in the online appendix⁹.

In Fig. 9, in the subplots on the top and middle rows, the *x* axis refers to time and the *y* axis refers to the visited model elements (AOIs). In the top row (a1–c1), each dot

Table 5 List of observations grouped by analytical direction

Observation	Label
5.2.1 Comprehension Phase —Addresses RQ3	
Observation 1	Only Essential Complexity Yields Demanding Discovery Phase.
Observation 2	Complexity (Essential and Accidental) Yields Demanding Post-Discovery Phase.
Observation 3	Complexity (Essential and Accidental) Yields More Revisits During Post-Discovery than Discovery.
Observation 4	Accidental Complexity Yields More Revisits in the Post-Discovery Phase than Essential Complexity.
5.2.2 Task-Relevant Versus Non-Task-Relevant Analysis —Addresses RQ4	
Observation 5	Both Task-Relevant Parts and Non-Task-Relevant Parts (Noise) Contribute to Cognitive Load.
Observation 6	Non-Task-Relevant Parts (Noise) Contribute More to Cognitive Load in Simple Tasks than in Complex Tasks.
5.2.3 Discovery Versus Post-Discovery Analysis —Addresses RQ5	
Observation 7	Non-Task-Relevant Parts (Noise) Contribute More to Cognitive Load in Simple Tasks than in Complex Tasks.
Observation 8	Non-Task-Relevant Parts (Noise) Contribute to Cognitive Load in Both Phases, but Unevenly.
Observation 9	Non-Task-Relevant Parts (Noise) Have a Decreased Contribution to Cognitive Load in Post-Discovery.

⁹Online appendix: <https://doi.org/10.5281/zenodo.17991805>

represents a fixation on an AOI, with vertical placement reflecting visit order. The black vertical lines mark the first visit to each of the two AOIs referenced in the task assertion. This allows us to distinguish between two phases: one before the second vertical line, which we call the *discovery phase* and one after that line, which we call the *post-discovery phase*. In other words, the discovery phase corresponds to the period when the AOIs referred to in the task assertion are not yet visited, while the post-discovery phase corresponds to the period when both AOIs referred to in the task assertion have been visited.

The middle row (a2–c2) adds an envelope curve (dash-dotted) and slope approximations (dashed lines), indicating the rate of exploration. A steep slope in early task phases suggests active scanning, followed by a plateau in the post-discovery phase, where users revisit already explored AOIs. This transition is a general trend occurring across several participants.

In the bottom row (a3–c3), the x axis still refers to the time and the y axis refers to the fixation duration, in milliseconds (ms). In the graphs, the durations are visualized as follows: Light gray: short fixations (< 250ms); Gray: mental processing fixations (250–500ms) and Black: deep mental processing fixations (> 500ms).

Observation 1: only essential complexity yields demanding discovery phase Comparing tasks on simple models to those on models with essential complexity, we observe longer discovery times in the latter (e.g., 14.79s in a2 vs. 79.08s in b2 for P22). This is consistent across participants ($M_{simple} = 18.96$, $M_{ess} = 54.12$, $d > 1$, $p < .001$). Additionally, tasks on models with essential complexity result in more deep mental processing fixations during discovery (e.g., 1 in a3 vs. 8 in b3), again statistically confirmed ($M_{simple} = 1.347$, $M_{ess} = 3.429$, $d > .5$, $p < .001$).

For accidental complexity, the effect is less pronounced. Discovery duration differs modestly (14.79s in a2 vs. 20.74s in c2 for P22) with medium practical significance ($d > .5$), but this is not statistically significant across participants ($p = 0.114$). Similarly, the number of deep processing fixations is comparable (1 in a3 vs. 0 in c3), with medium practical significance ($d > .5$) and no significant difference ($p = 0.373$).

Observation 2: complexity (essential and accidental) yields demanding post-discovery phase Post-discovery durations are longer in tasks on models with high essential or accidental complexity compared to tasks on simple models: 137.18s and 102.16s in b2 and c2 (for high essential and accidental complexity respectively), compared to 24.49s in a2 (low essential and accidental complexity). This trend is general ($M_{simple} = 25.51$, $M_{ess} = 90.89$, $M_{acc} = 80.74$, with $d > 1$ and $p < .001$ for both).

Similarly, more deep mental processing fixations occur in tasks on models with high essential or accidental complexity (e.g., 15 in b3, 11 in c3, vs. 7 in a3), again supported by broader data ($M_{simple} = 2.816$, $M_{ess} = 8.980$, $M_{acc} = 9.755$, all $d > 1$ and $p < .001$).

Observation 3: complexity (essential and accidental) yields more revisits during post-discovery than discovery Participants tend to revisit AOIs more during post-discovery than discovery phases. For P22, revisit ratios increase from discovery to post-discovery (i.e., 0.619 to 0.780 for a task on a model with essential complexity). The observation is statistically supported for tasks on models with high essential complexity ($M_{disc.} = 0.606$, $M_{post-disc.} = 0.714$, $d > 1$, $p < 0.01$) and high accidental complexity

($M_{disc.} = 0.585$, $M_{post-disc.} = 0.778$, $d > 1$, $p < 0.01$). However, an exception occurs for tasks on simple models where practical significance remains small ($d = 0.256 > 0.2$), and the $p = 0.209 > 0.05$, indicating no significant difference in revisit ratio between the discovery and post-discovery phases.

Observation 4: accidental complexity yields more revisits in the post-discovery phase than essential complexity By comparing the revisit ratios on models with essential and accidental complexity in the post-discovery phase, it emerged that accidental complexity yields more revisits in the post-discovery phase than essential complexity. For P22, this comparison is valid with $\rho_{acc} = 0.864 > 0.780 = \rho_{ess}$. This observation is statistically supported for the other participants ($M_{acc} = 0.778$, $M_{ess} = 0.714$, $d > .5$, $p < .001$).

Task-relevant versus non-task-relevant analysis

This part of the analysis addresses **RQ4**: *Does the complexity of the non-task-relevant parts of process models affect measures of users' cognitive load during comprehension tasks?*

To investigate this, we analyze data from **Set B**. Depending on the task, the task-relevant and non-task-relevant parts of a model may correspond either to the simple or to the complex part of the composed model. In what follows, we refer to the non-task-relevant part of a model as *noise*.

Table 6 presents two sets of results for two cognitive load indicators: AOI run count (cognitive integration) and the number of mental processing fixations. For both essential and accidental complexity, the table reports: (1) Descriptive statistics (M_{Noise} and $M_{Rel.}$, representing the mean values of the cognitive load measures on the noise and task-relevant parts of the model, respectively), along with the corresponding p -values from pairwise comparisons between these two model regions; (2) Difference scores (Δ), representing the overall mean of per-task, per-participant differences in proportions of

Table 6 Descriptive and inferential statistics. Abbreviations: #: number of observations; Noise: non-task-relevant part of a model; Rel.: task-relevant part of a model; M_{Noise} and $M_{Rel.}$: calculated means for measures on the noise and the task-relevant part of a model, respectively; ES: effect size using *Cohen's d*. ES can be categorized into small ($= 0.2$), medium ($= 0.5$), or large (≥ 0.8). A p -value $< .05$ means that the results of pairwise comparisons are significant; Δ : overall mean of difference of proportions between the noise and the task-relevant part

Complexity type	Task type	Statistics									
		Noise vs Rel.						Complex vs Simple			
		#	MNoise	MRel.	ES: d	p-value	#	Δ (%)	ES: d	p-value	
AOI Run Count											
Essential	Complex	50	16.72	113.4	> 1	< 0.01	50	73.82	> 1	< 0.01	
	Simple	49	30.78	48.98	> 1	< 0.01		39.72			
Accidental	Complex	49	17.45	86.2	> 1	< 0.01	48	61.88	> .5	< 0.05	
	Simple	49	18.98	48.08	> .5	< 0.01		50.54			
Number of Mental Processing Fixations											
Essential	Complex	42	3.7	24.37	> 1	< 0.01	35	60.94	> 1	< 0.01	
	Simple	38	7.93	10.02	0.25	< 0.05		44.06			
Accidental	Complex	37	3.18	21.55	> 1	< 0.01	27	70.29	> 1	< 0.01	
	Simple	32	4.56	9.57	> .8	< 0.01		45.6			

cognitive load indicators across the task-relevant and noise parts, along with p -values comparing complex and simple tasks. This proportion-based view complements the raw values and helps identify broader trends across complexity types and task types.

After reviewing the data on deep mental processing fixations, we decided to omit this indicator from the fine-grained analysis. Although some differences align with our expectations and are statistically supported, sample sizes for deep mental processing fixations are substantially low—in some cases reaching zero.

Observation 5: both task-relevant parts and non-task-relevant parts (noise) contribute to cognitive load Across all task types, the difference in proportions between task-relevant and non-task-relevant parts (Δ , in Table 6) is consistently below 100%. This shows that noise also contributes to users' cognitive load in terms of cognitive integration and mental processing. This holds true regardless of the complexity type.

Observation 6: non-task-relevant parts (noise) contribute more to cognitive load in simple tasks than in complex tasks We observe that the difference Δ is consistently lower for simple tasks than for complex tasks, indicating a greater relative contribution from noise during simple tasks. This is statistically supported by the p -values comparing Δ across task types (complex vs. simple), all of which are below 0.05. The effect sizes are globally medium to large, confirming the practical relevance of these differences. An exception occurs for the number of mental processing fixations in simple tasks with essential complexity, where the practical significance is small ($d = 0.25 > .2$) despite a statistically significant result ($p < 0.05$). Overall, the influence of noise is significantly higher in simple tasks, regardless of whether complexity is essential or accidental.

Discovery versus post-discovery analysis

This part of the analysis addresses **RQ5: How does the complexity of the non-task-relevant parts of process models affect users' cognitive load during different phases of comprehension tasks?**

As RQ5 is an extension of RQ4, this analysis also focuses on composed models that contain parts with varying levels of complexity (i.e., **Set B**).

Figure 10 visualizes how mental processing fixations are distributed (as proportions) across task phases—discovery and post-discovery—and task types—simple and

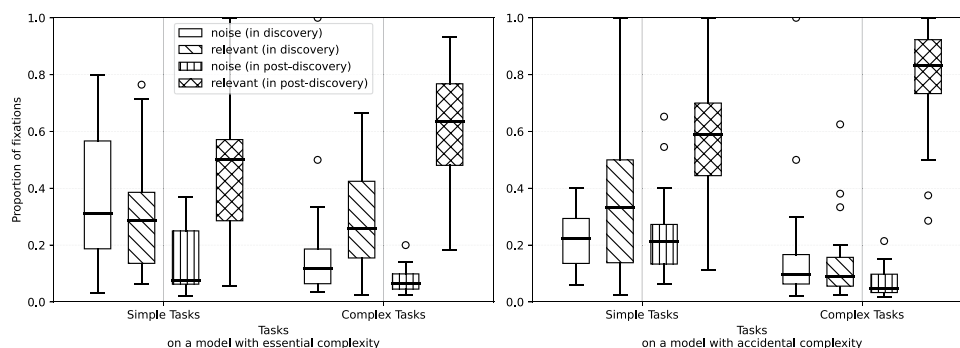


Fig. 10 Proportions of mental processing fixations in tasks with essential complexity (left) and accidental complexity (right). Boxplots show distributions of proportions by task type (simple or complex) and task phase (discovery or post-discovery)

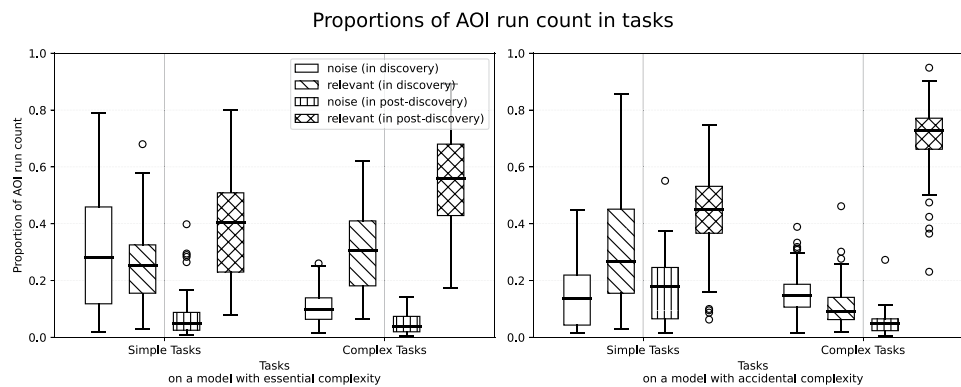


Fig. 11 Proportions of AOI run count in tasks with essential complexity (left) and accidental complexity (right). Boxplots show distributions of proportions by task type (simple or complex) and task phase (discovery or post-discovery)

complex. Figure 11 shows the same structure for AOI run count (i.e., cognitive integration). The description of Fig. 10 applies analogously to Fig. 11.

Each figure is divided into two subplots: one for tasks involving essential complexity (left) and one for accidental complexity (right). For each task type, the proportions are grouped into four boxplots: two showing noise and task-relevant contributions during the discovery phase, and two for the post-discovery phase.

As previously discussed in Sect. [Discovery Versus Post-Discovery Analysis](#), the absence of results for deep mental processing fixations stems from their very low (and in some cases null) sample sizes.

Observation 7: non-task-relevant parts (noise) contribute more to cognitive load in simple tasks than in complex tasks Across both measures (mental processing fixations and AOI run count), the median values of noise-related boxplots (i.e., boxplots without hatching and with vertical hatching) are generally higher for simple tasks than for complex tasks (except for AOI run count for accidental complexity). This confirms earlier findings (Sect. [Discovery Versus Post-Discovery Analysis](#)) that noise contributes more strongly to mental processing and cognitive integration in simple tasks compared to complex tasks.

Observation 8: non-task-relevant parts (noise) contribute to cognitive load in both phases, but unevenly Noise-related fixations occur throughout the task, as shown by their presence in both the discovery and post-discovery phases. However, the distribution is not uniform: noise is clearly more prominent during the discovery phase. This is particularly visible in the mental processing fixation plots, where boxplots representing noise (no hatching) often overlap those for task-relevant fixations (with diagonal hatching), suggesting that the differences may not be statistically significant. An exception occurs in complex tasks with essential complexity, where task-relevant fixations are clearly dominant even during discovery.

Observation 9: non-task-relevant parts (noise) have a decreased contribution to cognitive load in post-discovery In the post-discovery phase, a marked separation appears between task-relevant and noise boxplots. Boxplots for task-relevant areas (with cross-hatching) show higher medians than those for noise (with vertical hatching), indi-

cating that attention shifts toward task-relevant areas once the model has been partially explored. This pattern holds across both indicators and both complexity types and is particularly pronounced in complex tasks.

This time-based perspective adds granularity to the findings of **RQ4**. It shows that while noise contributes to cognitive effort throughout the task, it plays a more significant role early on—especially in simple tasks—and diminishes in the later stages as users focus on task-relevant areas. This highlights the exploratory nature of the discovery phase and confirms the reduced impact of noise in post-discovery.

Discussion

The discussion first focuses on the results from the coarse-grained analysis, followed by the results from the fine-grained analysis. Subsequently, the discussion continues with the presentation of our study's implications, before addressing potential threats to validity.

Coarse-grained analysis

To address RQ1, we performed a coarse-grained analysis grounded in a research model that investigates the impact of model complexity on cognitive load. Using a comprehensive suite of metrics, each designed to capture specific aspects of essential and accidental complexity, the results demonstrate two key findings: (1) When the complexity of two process models is clearly distinguishable across all selected metrics, we can expect correspondingly distinct estimates of cognitive load. (2) The direction of specific complexity metrics (i.e., whether a model is “more” or “less” complex) consistently aligns with selected indicators of cognitive load (i.e., higher complexity correlates with higher load).

In summary, these findings show that when a comprehensive set of complexity metrics indicates a significant difference between two models, this distinction is also reflected in the cognitive load of users during tasks involving these models.

This result contrasts with previous findings in empirical software engineering, where no significant relationship was found between complexity and cognitive load (Couceiro et al. 2019), or could only be established to a very limited degree (Peitek 2022; Hao et al. 2023).

To address RQ2, we performed a complementary coarse-grained analysis based on a research model that incorporates task complexity as a factor that affects the relationship between model complexity and cognitive load.

Our results indicate: (1) When two tasks are clearly distinct in complexity, that is, they target model regions with varying complexity measures, differences in cognitive load are again observable. (2) The selected cognitive load indicators follow a trend that aligns with the complexity of the parts of the model addressed by the respective tasks. A minor exception concerns comprehension accuracy, which seems to reach a practical limit in our composed models. This can be explained by the fact that, under accidental complexity, both model parts depict the same process and differ only in their graphical representation, leading to small practical differences despite visual variations.

These findings show that task characteristics significantly impact the observed relationship between process model complexity and cognitive load, qualifying the answer to RQ2. Thus, both the structure of the model and the task jointly influence the cognitive demands.

The relationship between complexity metrics and cognitive load has been relatively underexplored in the process modeling literature, with few exceptions (Abbad-Andaloussi et al. 2023). Previous studies have more commonly examined how the complexity of the process model relates to user performance (e.g., in terms of comprehension or error rates) and perceived difficulty (Mendling 2007; Figl and Laue 2011; Mendling et al. 2012; Burattin et al. 2017). In contrast, our work provides additional and stronger empirical evidence that both essential and accidental complexity metrics correlate with cognitive load. This is achieved using a broader set of indicators, including behavioral and physiological responses, thus offering a deeper understanding of the underlying cognitive processes. Furthermore, our results clearly demonstrate that task characteristics play a critical role in shaping the relationship between complexity metrics and cognitive load. This finding underscores the limitation of relying solely on complexity metrics for predicting cognitive load as such predictions risk being incomplete or misleading if the task characteristics are not taken into account.

Fine-grained analysis

The second part of our analysis builds upon the previously established relationships and shifts the focus towards fine-grained explorations of user behavior and cognitive load.

To explore RQ3, we performed an inductive analysis of visual behavior over time. We introduce the concept of the task envelope as a coarse temporal boundary that divides task execution into two distinct phases: *discovery* and *post-discovery*. The discovery phase shows by short duration, low revisit ratios, and relatively low cognitive load—suggesting an initial exploration or “search” stage (Wolfe 1994). The post-discovery phase is longer, with higher revisit ratios and increased indicators of deep mental processing (e.g., long fixations), similar to an “inference” phase (Kim et al. 2000) in which participants cognitively manipulate and relate the discovered information.

The distinction between these phases is consistent with prior findings on how process model comprehension evolves from perceptual to semantic reasoning. (Reijers and Mendling 2011) observed that understanding develops as users shift from syntactic parsing of model elements toward semantic integration of their relations, reflecting a transition from surface-level to deeper cognitive processing. Similarly, Pinggera et al. (2012) identified alternating cognitive modes during process modeling, ranging from exploratory to analytical behaviors, which parallel the dynamics observed here between discovery and post-discovery. Taken together, these results align with the information-foraging and processing phases described in sensemaking practices (Chen et al. 2025), clarifying how cognitive processes unfold temporally during interaction with process models.

Additionally, by comparing models with predominantly accidental versus essential complexity, we found that models with more layout-induced (accidental) complexity tend to elicit higher revisit ratios in the post-discovery phase, suggesting a more effortful mental reorganization process. This suggestion aligns with the outcomes of the research in (Figl et al. 2024) emphasizing the negative cognitive effects of bad model layout emerging from high accidental complexity.

To answer RQ4 and RQ5, we introduced the notion of “noise”—defined as the non-task-relevant parts of the process model. For RQ4, we compared the cognitive impact of task-relevant versus non-task-relevant (noise) regions. Our results show that both

contribute to cognitive load, but noise has a proportionally greater effect in simpler tasks, likely due to the lower cognitive demand of the task itself, making distractions more impactful. We observed a minor exception for the number of mental processing fixations in simple tasks with essential complexity. In this configuration, the noise corresponds to the complex part of the model, which is typically skimmed during an initial *discovery* phase and thus requires only limited mental effort. Conversely, the relevant part corresponds to the simple portion of the model, where mental processing demands are also expected to be low. As a result, only a few mental processing fixations (as manifestations of cognitive load) occur in both regions, leading to a low practical significance. This pattern is consistent with other results showing reduced cognitive load on simple models and less intensive visual behavior during the *discovery* phase, suggesting that it reflects a general phenomenon rather than an anomaly. Overall, these findings confirm that task-relevant and non-task-relevant regions both contribute to cognitive load, with noise having a proportionally greater impact in simple tasks.

For RQ5, we examined how the effect of noise evolves over time. We found that noise does not affect users uniformly throughout the task: In the discovery phase, noise plays a more prominent role. In the post-discovery phase, its effect diminishes as users shift their attention to the relevant parts of the model. These findings not only confirm that noise contributes to cognitive load, but also highlight when and how this impact occurs. Moreover, they advance existing research that assigns cognitive load to task-relevant information (Petrusel and Mendling 2013) by demonstrating that the characteristics of the non-task-relevant information also influence cognitive load, particularly in the early stages of comprehension.

Finally, our observations suggest that most visual attention in the post-discovery phase is focused on task-relevant areas, supporting the interpretation that this phase corresponds to goal-directed inference rather than general exploration.

Implications

This study refines and extends existing knowledge on how process model complexity relates to cognitive load by revealing the joint influence of model- and task-related factors. Whereas prior work primarily focused on static model characteristics Mendling 2007; Bernstein and Soffer 2015; Burattin et al. 2017, our results demonstrate that the complexity of the *task* addressing a model region significantly modulates users' cognitive effort. This challenges the implicit assumption in many BPM complexity studies that model complexity alone can predict cognitive load, suggesting instead that comprehension depends on an interaction between model structure and task demands.

Furthermore, by uncovering distinct 'discovery' and 'post-discovery' gaze phases, our findings complement the dual-phase cognitive processing models proposed in sense-making theory (Chen et al. 2025). These results indicate that process model comprehension involves an initial phase of exploration and orientation followed by a phase of reasoning and integration—each exhibiting different patterns of response to complexity and task relevance.

From a practical standpoint, our findings encourage BPM practitioners and tool developers to move beyond global complexity assessments. Designers should consider how task scope and spatial focus shape users' cognitive load and employ adaptive

visualization mechanisms to minimize the impact of non-task-relevant elements (i.e., noise), especially during early discovery phases.

Methodologically, this study advances the empirical toolkit for BPM research by demonstrating how to integrate a cognitive approach into the study of process model comprehension. By leveraging eye-tracking, our approach combines behavioral and physiological indicators of cognitive load with temporal segmentation, enabling a more precise mapping of cognitive load dynamics over time. The experimental design proposed in this study provides a replicable framework for future research seeking to disentangle essential from accidental complexity while balancing experimental control and ecological validity.

Threats to validity

Our work can be subject to internal, ecological and external validity threats presented in the following paragraphs.

Internal validity

Two primary concerns may affect internal validity: (1) Whether the models used were sufficiently realistic, and (2) Whether cognitive load was accurately captured.

To address (1), we followed an incremental modeling strategy. Models were developed progressively—from basic ones with clear complexity differences (used in prior studies), to more advanced ones containing parts with varying complexity, better reflecting realistic conditions.

Regarding (2), we took several measures to mitigate confounding factors (see [Section Research Method](#)). These measures include creating a limited number of tasks to reduce potential learning effects due to repeated exposure to the models, and randomizing the tasks among participants. In addition, we controlled the experimental setup and provided all participants with a standardized introduction to BPMN and the task structure. Finally, we measured cognitive load using multiple complementary indicators, including subjective, performance, physiological and behavioral measures that previous studies have successfully used (Abbad-Andaloussi et al. [2023](#)). Together, these design choices enhance confidence in the validity of our findings.

Ecological validity

The use of non-realistic activity labeling (i.e., random two-letter codes) in the modeling design of our material had a potential impact on ecological validity, as the results might not fully hold for domain-specific or semantically rich models. To address this threat, we acknowledged that anonymized, semantically reduced activity labels could limit the ecological validity of our findings, since process models in real-world settings typically contain domain-specific terminology that supports comprehension through semantic cues.

In addition, the study's focus on binary assertion verification tasks—although advantageous for ensuring experimental control and comparability—further constrains generalizability, as it does not capture the diversity of modeling-related activities that process models support in practice, such as explanation, problem-solving, or redesign.

Nevertheless, our primary goal was to isolate the effects of structural complexity on cognitive load while minimizing confounding influences from domain familiarity, semantic richness, and task variability. By controlling for these factors, we ensured that

the observed differences in participants' cognitive effort stemmed primarily from the graphical characteristics and inherent structural content of the models rather than from textual or task-related artifacts. Consequently, while our findings may not fully generalize to domain-specific or task-diverse modeling contexts, they provided a controlled basis for understanding how visual and structural features contribute to cognitive load independently of semantic knowledge and task complexity.

External validity

The relatively small sample size may limit generalizability. However, this is consistent with previous studies that involve eye tracking in similar domains (Petrusel and Mendling 2013; Peitek 2022; Hao et al. 2023), reflecting the methodological complexity of such experiments (Holmqvist et al. 2011). To enhance transferability, we based all models on well-known control-flow BPMN patterns (Muehlen and Recker 2008), which improves comparability and applicability to broader contexts.

Conclusion and future work

This study extends our research presented in (Sorg et al. 2025). Building upon our previous investigations, it further explores the relationship between complexity metrics and users' cognitive load by introducing the task as a factor affecting this relationship. Additionally, the investigation of user visual behavior is expanded through the implementation of descriptive and inferential statistics, thereby substantiating our findings. Our multi-granular research method allows a comprehensive analysis of the different aspects of the relationship between process model complexity and cognitive load. At a coarse-grained level, the findings demonstrate first that an extensive set of essential and accidental complexity metrics can align well with cognitive load during comprehension tasks on simple and complex process models. However, in a complementary manner, the findings also clearly demonstrate the influence of the task on the relationship between complexity and cognitive load. At a fine-grained level, the exploration of users' visual behavior over time allowed the identification of two comprehension phases (i.e., discovery and post-discovery) which relate to information search and inference and underlie different levels of cognitive processing. Moreover, our analysis showed that task-relevant and non-task-relevant parts of process models contribute to cognitive load differently during both phases.

As future work, it is important to delve deeper into users' cognitive processes, particularly those underlying information search and inference, as well as mental reorganization and cognitive integration, which are likely elicited by essential and accidental process model complexity. A deeper understanding of these mechanisms would ultimately support the anticipation of comprehension problems, the reduction of cognitive load, and the prevention of modeling errors. An important direction to advance this line of research is to extend the investigation beyond binary assertion verification and include more elaborated tasks—such as explanation, problem-solving, or model redesign—that better reflect how process models are used in practice and strengthen the ecological validity of future studies. In this regard, exploring how process model complexity influences users' performance and cognitive effort appears promising, as both characteristics reflect how objective complexity translates into felt difficulty and cognitive demand during comprehension. Attention should also be given to user characteristics (e.g., level of

process modeling expertise) and their interaction with complexity and visual behavior across search and inference phases, as these factors are likely to modulate the mechanisms of comprehension and cognitive load regulation. Extending this research to expert modelers and real-world business contexts would represent a natural progression of this line of research, allowing practical implementation and validation of these mechanisms in professional modeling environments. In support of the aforementioned research, the use of eye-tracking technology offers opportunities to develop novel methodological approaches. Future studies could leverage eye-tracking data through machine learning techniques to identify complex patterns of cognitive processing. Combining these developments may lead to adaptive systems capable of automatically anticipating error-prone situations by continuously analyzing users' visual behavior with greater accuracy.

Author contributions

Th.S.: wrote the main manuscript text, and prepared all figures and tables. A.A.: reviewed and edited the manuscript. E.K.: reviewed the manuscript. B.W.: reviewed and edited the manuscript.

Funding

The authors did not receive support from any organization for the submitted work.

Data availability

Data is provided within the manuscript and in the online appendix in the following repository: <https://doi.org/10.5281/zenodo.17991805>.

Declarations

Consent to participate

The study participants provided their informed consent to contribute with their responses and to have their eye-tracking and video/audio data recorded. Participation was voluntary, and all participants were informed of their right to withdraw at any time without penalty. The collected data was anonymized to ensure the privacy of the participants and used solely for the purposes of this study.

Competing interests

The authors declare no competing interests.

Received: 25 July 2025 / Accepted: 9 December 2025

Published online: 24 January 2026

References

- Abbad-Andaloussi A (2021) A framework for enhancing the modeling and comprehension of declarative process models. PhD thesis, Technical University of Denmark. <https://www.alexandria.unisg.ch/handle/20.500.14171/110801>,
- Abbad-Andaloussi A (2023) On the relationship between source-code metrics and cognitive load: A systematic tertiary review. *J Syst and Softw* 198:111619. <https://doi.org/10.1016/j.jss.2023.111619>
- Abbad-Andaloussi A, Burattin A, Slaats T, Kindler E, Weber B (2023) Complexity in declarative process models: Metrics and multi-modal assessment of cognitive load. *Expert Syst with Appl*. <https://doi.org/10.1016/j.eswa.2023.120924>
- Abbad-Andaloussi A, Franceschetti M, López HA, Schreiber C, Weber B (2025) When a model gives you mixed signals: cognitive effects and visual behavior. *Process Sci* 2(1). <https://doi.org/10.1007/s44311-025-00022-8>
- Abbad-Andaloussi A, Lübke D, Weber B (2023) Conducting eye-tracking studies on large and interactive process models using eyemind. *SoftwareX* 24. Articleid:101564. <https://doi.org/10.1016/j.softx.2023.101564>
- Abbad-Andaloussi A, Schreiber C, Weber B (2024) On the Task-Specific Effects of Fragmentation in Modular Process Models. *Springer*, pp 18–37. https://doi.org/10.1007/978-3-031-75872-0_2
- Abbad-Andaloussi A, Zerbato F, Burattin A, Slaats T, Hildebrandt TT, Weber B (2020) Exploring how users engage with hybrid process artifacts based on declarative process models: a behavioral analysis based on eye-tracking and thinkaloud 20(5):1437–1464. <https://doi.org/10.1007/s10270-020-00811-8>
- Antinyan V (2020) Evaluating essential and accidental code complexity triggers by practitioners' perception. *IEEE Softw* 37(6):86–93. <https://doi.org/10.1109/ms.2020.2976072>
- Basili VR, Caldiera G, Rombach HD (1994) The goal question metric approach. *Encycl of Softw Eng* 528–532. John Wiley & Sons, (<https://api.semanticscholar.org/CorpusID:13884048>)
- Bera P, Soffer P, Parsons J (2019) Using eye tracking to expose cognitive processes in understanding conceptual models. *MIS Q* 43(4):1105–1126. <https://doi.org/10.25300/MISQ/2019/14163>
- Bernstein V, Soffer P (2015) Identifying and quantifying visual layout features of business process models. In: Gaaloul, K., Schmidt, R., Nurcan, S., Guerreiro, S., Ma, Q. (eds.) *Enterp, Bus-Process and Inf Syst Modeling* 200–213. Springer, Cham https://doi.org/10.1007/978-3-319-19237-6_13
- Burattin A, Bernstein V, Neurauter M, Soffer P, Weber B (2017) Detection and quantification of flow consistency in business process models. *Softw & Syst Modeling* 17(2):633–654. <https://doi.org/10.1007/s10270-017-0576-y>

- Chen F, Zhou J, Wang Y, Yu K, Arshad SZ, Khawaji A, Conway D (2016) Robust Multimodal Cognitive Load Measurement. *Hum-Comput Interact Ser*. Springer, Switzerland. <https://doi.org/10.1007/978-3-319-31700-7>
- Chen T, Weber B, Shanks G, Demartini G, Indulska M, Sadiq S (2025) How Do Experts Make Sense of Integrated Process Models?. Springer 3–19. https://doi.org/10.1007/978-3-031-94571-7_1
- Cheng C-Y (2008) Complexity and Usability Models for Business Process Analysis. The Pennsylvania State University
- Couceiro R, Carvalho P, Branco MC, Madeira H, Barbosa R, Duraes J, Duarte G, Castelhan J, Duarte C, Teixeira C, Laranjeiro N, Medeiros J (2019) Spotting problematic code lines using nonintrusive programmers' biofeedback. In 2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE), IEEE Comput Soc Press, Los Alamitos, CA, USA, 93–103. <https://doi.org/10.1109/issre.2019.00019>
- Djurica D, Kummer T-F, Mendling J, Figl K (2025) Investigating the impact of representation features on decision model comprehension. *Eur J Inf Syst* 34(1):21–45. <https://doi.org/10.1080/0960085x.2023.2270456>
- Dobesova Z, Malcik M (2015) Workflow diagrams and pupil dilatation in eye-tracking testing. In 2015 13th International Conference on Emerging eLearning Technologies and Applications (ICETA), IEEE, pp. 1–6. <https://doi.org/10.1109/iceta.2015.7558460>
- Duchowski AT, Krejtz K, Gehrer NA, Bafna T, Baekgaard P (2020) The low/high index of pupillary activity. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI'20, ACM, New York, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376394>
- Dumas M, La Rosa M, Mendling J, Reijers HA (2013) Fundamentals of Business Process Management. Springer. <https://doi.org/10.1007/978-3-642-33143-5>
- Fahland D, Lübke D, Mendling J, Reijers H, Weber B, Weidlich M, Zugal S (2009) Declarative versus Imperative Process Modeling Languages: The Issue of Understandability, vol 29. Springer, Berlin, Heidelberg, pp 353–366. https://doi.org/10.1007/978-3-642-01862-6_29
- Figl K, Laue R (2011) Cognitive complexity in business process modeling. *Adv Inf Syst Eng* 452–466. Springer, (https://doi.org/10.1007/978-3-642-21640-4_34)
- Figl K, Laue R (2015) Influence factors for local comprehensibility of process models. *Int J Hum-Comput Stud* 82:96–110. <https://doi.org/10.1016/j.jihcs.2015.05.007>
- Figl K, Soffer P, Weber B (2024) Guiding attention in flow-based conceptual models through consistent flow and pattern visibility. *Decis Support Syst* 185:114292. <https://doi.org/10.1016/j.dss.2024.114292>
- Findik-Coşkunçay D, Kırık MP (2022) An investigation of the relationship between joint visual attention and product quality in collaborative business process modeling: a dual eye-tracking study. *Softw and Syst Modeling* 21(6):2429–2460. <https://doi.org/10.1007/s10270-022-00974-6>
- Franceschetti M, Abbad-Andaloussi A, Schreiber C, López HA, Weber B (2024) Exploring the Cognitive Effects of Ambiguity in Process Models. Springer, pp 493–510. https://doi.org/10.1007/978-3-031-70396-6_28
- Gidlöf K, Wallin A, Dewhurst R, Holmqvist K (2013) Using eye tracking to trace a cognitive process: Gaze behaviour during decision making in a natural environment. *J Eye Mov Res* 6(1). <https://doi.org/10.16910/jemr.6.1.3>
- Glöckner A, Herbold A-K (2008) Information processing in decisions under risk: Evidence for compensatory strategies based on automatic processes. *SSRN Electron J* 42. <https://doi.org/10.2139/ssrn.1307664>
- Goldstine HH, Neumann J (1947) Planning and coding of problems for an electronic computing instrument. *Math and Logical Aspects of an Electron Comput Instrument*
- Hao G, Hijazi H, Durães J, Medeiros J, Couceiro R, Lam CT, Teixeira C, Castelhan J, Castelo Branco M, Carvalho P, Madeira H (2023) On the accuracy of code complexity metrics: A neuroscience-based guideline for improvement. *Front Neurosci* 16. <https://doi.org/10.3389/fnins.2022.1065366>
- Holmqvist K, Nyström M, Andersson R, Dewhurst R, Halszka J, Weijer J (2011) Eye Tracking : A Comprehensive Guide to Methods and Measures. Oxford University Press, Oxford, United Kingdom. <http://ukcatalogue.oup.com/product/9780199697083.do>
- Just MA, Carpenter PA. (1980) A theory of reading: From eye fixations to comprehension. *Psychological Rev* 87(4):329
- Kim J, Hahn J, Hahn H (2000) How do we understand a system with (so) many diagrams? cognitive integration processes in diagrammatic reasoning. *Inf Syst Res* 11(3):284–303. <https://doi.org/10.1287/isre.11.3.284.12206>
- Malinova M, Mendling J (2021) Cognitive diagram understanding and task performance in systems analysis and design. *MIS Q* 45(4):2101–2158. <https://doi.org/10.25300/misq/2021/15262>
- Mendling J (2007) Detection and prediction of errors in EPC business process models. PhD thesis, Wirtschaftsuniversität Wien.
- Mendling J (2008 April) Detection and prediction of errors in epc business process models. In: D. Wagner (ed.) *Ausgezeichnete Informatikdissertationen 2007*. Springer, Dagstuhl, Germany, Germany, pp 211–218.
- Mendling J, Reijers HA, Aalst WMP (2010) Seven process modeling guidelines (7pmg). *Inf and Softw Technol* 52(2):127–136. <https://doi.org/10.1016/j.infsof.2009.08.004>
- Mendling J, Reijers HA, Cardoso J (2008) What makes process models understandable? In: Alonso G, Dadam P, Rosemann M (eds.) *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg pp 48–63. https://doi.org/10.1007/978-3-540-75183-0_4
- Mendling J, Sánchez-González L, García F, Rosa ML (2012) Thresholds for error probability measures of business process models. *J Syst and Softw* 85(5):1188–1197. <https://doi.org/10.1016/j.jss.2012.01.017>
- Moody D (2009) The “physics” of notations: Toward a scientific basis for constructing visual notations in software engineering. *IEEE Trans on Softw Eng* 35(6):756–779. <https://doi.org/10.1109/tse.2009.67>
- Muehlen MZ, Recker J (2008) How Much Language Is Enough? Theoretical and Practical Use of the Business Process Modeling Notation. Springer, pp 465–479. https://doi.org/10.1007/978-3-540-69534-9_35
- Peitek N (2022) A neuro-cognitive perspective of program comprehension. phdthesis, Technischen Universität Chemnitz. <https://nbn-resolving.org/urn:nbn:de:bsz:ch1-qucosa2-790021> Accessed 2023-12-01.
- Petrusel R, Mendling J (2013) Eye-tracking the factors of process model comprehension tasks. In: Salinesi, C., Norrie, M.C., Pastor, Ó. (eds.) *Adv Inf Syst Eng* 224–239. Springer, Berlin, Heidelberg https://doi.org/10.1007/978-3-642-38709-8_15
- Petrusel R, Mendling J, Reijers HA (2017) How visual cognition influences process model comprehension. *Decis Support Syst* 96:1–16. <https://doi.org/10.1016/j.dss.2017.01.005>
- Peysakhovich V, Causse M, Scannella S, Dehais F (2015) Frequency analysis of a task-evoked pupillary response: Luminance-independent measure of mental effort. *Int J Psychophysiol* 97(1):30–37. <https://doi.org/10.1016/j.jippsycho.2015.04.019>
- Pichler P, Weber B, Zugal S, Pinggera J, Mendling J, Reijers HA (2011) Imperative versus Declarative Process Modeling Languages: An Empirical Investigation. Springer, pp 383–394. https://doi.org/10.1007/978-3-642-28108-2_37

- Pinggera J, Zugal S, Weidlich M, Fahland D, Weber B, Mendling J, Reijers HA (2012) Tracing the Process of Process Modeling with Modeling Phase Diagrams. Springer, Berlin, Heidelberg, pp 370–382. https://doi.org/10.1007/978-3-642-28108-2_36
- Reijers HA, Mendling J (2011) A study into the factors that influence the understandability of business process models. *IEEE Trans on Syst, Man, and Cybern - Part A: Syst and Humans* 41(3):449–462. <https://doi.org/10.1109/TSMCA.2010.2087017>
- Ritchi H, Jans M, Mendling J, Reijers HA (2019) The influence of business process representation on performance of different task types. *J Inf Syst* 34(1):167–194. <https://publications.aaahq.org/jis/article-pdf/34/1/167/9424/i1558-7959-34-1-167.pdf>
- Sánchez-González L, García F, Mendling J, Ruiz F (2010) Quality assessment of business process models based on thresholds. In: Meersman, R., Dillon, T., Herrero, P. (eds.) *On the Move to Meaningful Internet Syst: OTM* 78–95. Springer, (2010). https://doi.org/10.1007/978-3-642-16934-2_9
- Schreiber C, Abbad-Andaloussi A, Weber B (2023) On the cognitive effects of abstraction and fragmentation in modularized process models. *Lect Notes in Comput Sci* 359–376. Springer, https://doi.org/10.1007/978-3-031-41620-0_21
- Schreiber C, Abbad-Andaloussi A, Weber B (2024) On the cognitive and behavioral effects of abstraction and fragmentation in modularized process models. *Inf Syst* 125:102424. <https://doi.org/10.1016/j.is.2024.102424>
- Sorg T, Abbad-Andaloussi A, Kindler E, Weber B (2025) Process model complexity metrics, cognitive load and visual behavior: A multigranular eye-tracking analysis. In: Guizzardi, R., Pufahl, L., Sturm, A., Aa, H. (eds.) *Enterp, Bus-Process and Inf Syst Modeling*. *Lect Notes in Bus Inf Process* 87–103. Springer, (<https://www.alexandria.unisg.ch/handle/20.500.14171/123387>)
- Sweller J (2010) Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educ Psychol Rev* 22(2):123–138. <https://doi.org/10.1007/s10648-010-9128-5>
- Sweller J, Ayres P, Kalyuga S (2011) *Cognitive Load Theory. Explorations in the Learning Sciences, Instructional Systems and Performance Technologies*. Springer. <https://doi.org/10.1007/978-1-4419-8126-4>
- Sweller J, Merrienboer JGG, Paas FGWC (1998) Cognitive architecture and instructional design 10(3):251–296. <https://doi.org/10.1023/a:1022193728205>
- Vessey I, Galletta D (1991) Cognitive fit: An empirical study of information acquisition. *Inf Syst Res* 2(1):63–84. <https://doi.org/10.1287/isre.2.1.63>
- Wang W (2019) *The Effect of Rule Linking on Business Process Model Understanding*. Springer, pp 42–59. https://doi.org/10.1007/978-3-030-11809-9_5
- Wang W, Chen T, Indulska M, Sadiq S, Weber B (2022) Business process and rule integration approaches—an empirical analysis of model understanding. *Inf Syst* 104:101901. <https://doi.org/10.1016/j.is.2021.101901>
- Weber B, Fischer T, Riedl R (2021) Brain and autonomic nervous system activity measurement in software engineering: a systematic literature review. *J Syst and Softw* 178. <https://doi.org/10.1016/j.jss.2021.110946>
- Winter M, Neumann H, Pryss R, Probst T, Reichert M (2023) Defining gaze patterns for process model literacy – exploring visual routines in process models with diverse mappings. *Expert Syst with Appl* 213:119217. <https://doi.org/10.1016/j.eswa.2022.119217>
- Winter M, Pryss R, Probst T, Bas J, Reichert M (2022) Measuring the cognitive complexity in the comprehension of modular process models. *IEEE Trans on Cognit and Dev Syst* 14(1):164–180. <https://doi.org/10.1109/tcds.2020.3032730>
- Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2024) *Experimentation in Software Engineering*, 2nd edn. Springer. <https://doi.org/10.1007/978-3-662-69306-3>
- Wolfe JM (1994) Guided search 2.0 a revised model of visual search. *Psychonomic Bull & Rev* 1(2):202–238. <https://doi.org/10.3758/bf03200774>
- Yan Z, Dijkman R, Grefen P (2017) Generating process model collections. *Softw and Syst Modeling* 16(4):979–995. <https://doi.org/10.1007/s10270-015-0497-6>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.