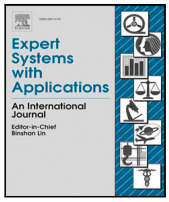
Expert Systems With Applications 233 (2023) 120924

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Complexity in declarative process models: Metrics and multi-modal assessment of cognitive load

Amine Abbad-Andaloussi [a,*], Andrea Burattin [b], Tijs Slaats [c], Ekkart Kindler [b], Barbara Weber [a]

[a] Institute of Computer Science, University of St. Gallen, 9000 St. Gallen, Switzerland
[b] Software Systems Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
[c] Department of Computer Science, University of Copenhagen, 2100 København, Denmark

ARTICLE INFO

Keywords:
Complexity metrics
Process model comprehension
Declarative process model
Cognitive load
Eye-tracking
Electrodermal activity

ABSTRACT

Complex process models can hinder the comprehension of the underlying business processes. While several metrics have been suggested in the literature to evaluate the complexity of imperative process models, little is known about their declarative counterparts. In this paper, we address this gap through a suite of metrics that we propose to capture the complexity of declarative process models. Following this, we empirically investigate the impact of complexity, as measured by the suggested metrics, on users' cognitive load when comprehending declarative process models. Therein, we use a multi-modal approach including eye-tracking and electrodermal activity. The findings of the empirical study provide evidence about the cognitive load emerging as a result of increased model complexity. Overall, the outcome of this paper presents empirically validated metrics to evaluate the complexity of declarative process models. Implementing these metrics and incorporating them in intelligent modeling tools would help assessing the complexity of declarative process models before being deployed. Furthermore, our empirical approach can be adopted by researchers in upcoming empirical studies to provide a multi-perspective assessment of users' cognitive load when engaging with process models.

1. Introduction

Process models provide a blueprint for system support and enable the communication between different stakeholders including domain experts and IT specialists (Dumas, La Rosa, Mendling, & Reijers, 2013). The modeling of processes is supported by formal languages, which need to be both machine-interpretable and comprehensible to humans. In the literature, these languages were categorized within the imperative-declarative paradigm (Fahland et al., 2009). Imperative languages represent all the paths of the process explicitly in the model. Declarative languages, in turn, emphasize the constraints governing the interplay between the process activities and represent the possible execution paths implicitly in the model (Fahland et al., 2009; Reichert & Weber, 2012). Imperative languages are typical for representing predefined and repetitive processes due to their sequence-flow nature (e.g., security screening in border control). Conversely, declarative languages allow representing flexible processes (e.g., patients diagnoses and treatments) concisely using their constraint-based approach.

The literature comprises a wide array of metrics to assess the complexity of process models, e.g., (Cardoso, Mendling, Neumann, & Reijers, 2006; Cheng, 2008; Gruhn & Laue, 2007; Latva-Koivisto, 2001; Mendling, 2007; Moreno-Montes de Oca & Snoeck, 2014; Polančič & Cegnar, 2017; Reijers, 2003; Reijers & Vanderfeesten, 2004; Sa, Garci, Ruiz, & Mendling, 2012). A notable example is Mendling's suite (Mendling, 2007) designed for imperative process models represented as Event-driven Process Chains (EPC) (Keller, Nüttgens, & Scheer, 1992). The suite is inspired by a large body of existing graph theory, software engineering and information theory metrics. It captures proprieties associated with the *size*, *cyclicity*, *concurrency*, *density*, *separability* and *connector heterogeneity* of process models (Mendling, 2007). The metrics of this suite can be easily applied to other imperative languages such as BPMN (Object Management Group (OMG), 2010) and Petri-nets (Petri, 1962) as they all share the same underlying paradigm. When it comes to declarative languages such as Declare (Pesic, Schonenberg, & van der Aalst, 2007) or Dynamic Condition Response (DCR) graphs (Hildebrandt & Mukkamala, 2011), little is known about the applicability of the existing metrics and their ability to capture the associated model proprieties. Looking at the representation of declarative process models, there are reasons to assume that some of the existing metrics need to be reformulated due to the implicit sequence-flow in declarative models and the possibility to have disconnected fragments within the same model. Motivated by this need,

* Corresponding author.
E-mail addresses: amine.abbad-andaloussi@unisg.ch (A. Abbad-Andaloussi), andbur@dtu.dk (A. Burattin), slaats@di.ku.dk (T. Slaats), ekki@dtu.dk (E. Kindler), barbara.weber@unisg.ch (B. Weber).

https://doi.org/10.1016/j.eswa.2023.120924
Received 20 March 2023; Received in revised form 26 May 2023; Accepted 28 June 2023
Available online 5 July 2023

the first research question aims at adapting the existing metrics to declarative process models. We define this question as follows: **RQ1. How to adapt the existing metrics to declarative process models?** To address this research question, we turn to the metrics proposed in Mendling (2007), study their adaptability and formulate new metrics that can apply to declarative languages.

The existing metrics capture a number of proprieties which have been associated with complexity and shown to challenge users when engaging with imperative process models during comprehension tasks (Figl & Laue, 2011; Reijers & Mendling, 2010; Sánchez-González, García, Mendling, & Ruiz, 2010). Following the cognitive load theory (Chen, Zhou, Wang, Yu, Arshad, Khawaji, et al., 2016; Paas, Tuovinen, Tabbers, & Van Gerven, 2003), this challenge is due to the inability of humans to cope with complex artifacts (e.g., process models) leading them to very high cognitive load and thus more difficulty to comprehend the process models at hand. While these insights have been verified for imperative process models (Figl & Laue, 2011; Reijers & Mendling, 2010; Sánchez-González et al., 2010), little is known about the declarative ones. To address this need, the second research question aims at evaluating the effects of the proprieties captured by the adapted metrics on users' cognitive load when engaging with declarative process models. Therein, we conduct an empirical study where we use the DCR notation as a proxy for declarative languages since it is supported by industry-grade modeling tools and has several use-case applications documented in the literature, e.g., Hildebrandt et al. (2020) and López, Debois, Hildebrandt, and Marquard (2018). Accordingly, we define the following research question: **RQ2. Do the proprieties captured by the newly adapted metrics affect users' cognitive load when engaging with declarative process models in DCR?** To answer this research question, we formulate a set of hypotheses where it is expected that complex declarative models, as estimated by each of our individual metrics, would yield a higher cognitive load because users are required to extract, recall and integrate more information into their mind. We test our hypotheses in a controlled experiment where participants' cognitive load is measured continuously along entire comprehension tasks on DCR models of different complexity levels. We use a multi-modal approach supported by eye-tracking (Holmqvist, Nyström, Andersson, Dewhurst, Jarodzka, & van de Weijer, 2011) and electrodermal activity (EDA) (Critchley, 2002) measures. To the best of our knowledge, our paper is the first to deploy a broad spectrum of multi-modal measures of cognitive load in the process modeling literature. Our findings provide empirical evidence showing that the newly adapted metrics are associated with users' cognitive load. As a result, our study is also the first to deliver empirically validated complexity metrics for declarative models.

The outcome of this work is twofold. On the one hand, it delivers a suite of complexity metrics serving to evaluate the quality of declarative process models and providing heuristics to guide the discovery of process models with reduced complexity from event-logs. Moreover, our empirical findings demonstrate the relevance of the proprieties measured by our metrics and thus lay the foundation for a new set of guidelines for declarative process models. On the other hand, with our multi-modal approach assessing cognitive load, a new class of experiments is made possible. Therein, one can quantify and compare users' cognitive load based on measurements derived from different modalities. Given the continuous nature of these measurements, our approach has also the potential to be used for more advanced analyses where it is possible to pinpoint exactly where — but also when cognitive load occurs. Additionally, with further development, it can be moved to online settings to provide intelligent modeling tools providing ad-hoc support to users based on real-time assessment of their cognitive load.

In the remainder of this paper, Sections 2 and 3 present the background and related work respectively. Section 4 presents the newly adapted metrics for declarative process models. Section 5 describes the empirical study validating these metrics. The findings of the empirical study are reported and discussed in Section 6, while the underlying limitations are presented in Section 7. Finally, the paper is concluded in Section 8.

2. Background

This section introduces the core concepts of declarative languages needed to define our complexity metrics (cf. Section 2.1), then introduces the DCR notation which is used as a proxy for declarative languages in our empirical study (cf. Section 2.2). Afterward, it provides a background on cognitive load (cf. Section 2.3) and presents a set of common measures capturing it (cf. Section 2.4).

2.1. Declarative languages

Process models are represented formally using imperative or declarative languages. As mentioned in Section 1, imperative languages specify all the paths of a process explicitly in the model, whereas declarative languages follow a constraint-based approach describing the interplay between the process activities without explicitly showing all the execution paths allowed in the process (Fahland et al., 2009; Reichert & Weber, 2012). As for the semantics, any execution path satisfying the given constraints is allowed by the model. This core feature of declarative languages allows them to represent flexible processes – with many execution paths – in a compact manner (Fahland et al., 2009; Reichert & Weber, 2012).

Declarative process models are typically represented as graphs composed of nodes and edges. The nodes refer to the activities of the process, while the edges refer to constraints prescribing the interplay between these activities. Activities that are not linked with constraints can be executed several times and at any point in time. Similarly, blocks of activities forming *a weakly connected component*[1] in the graph can be executed independently without being influenced by the other weakly connected components within the same graph. Hence, declarative process models can have disconnected fragments within the same model. As for the constraints, their semantics differ from one language to another. Declare, for instance, comprises 14 standard constraints allowing to model control-flow patterns associated with existence, choice, relation, negation and branching (Reichert & Weber, 2012). DCR, in turn, has 6 core relations allowing to model patterns including conditions, milestones, dynamic inclusions, dynamic exclusions, responses, and no-responses (cf. Section 2.2).

The (adapted) metrics which will be presented in Section 4 are based on these basic features and thus can apply to any declarative model with such characteristics. The models used for the empirical study, in turn, are based on DCR which we consider as a representative for declarative languages (cf. Section 2.2). As motivated in Section 1, we focus on DCR because of the availability of industry-grade tools supporting the modeling of DCR graphs (Marquard, Shahzad, & Slaats, 2016) and the increasing number of applications documented in the literature (Hildebrandt et al., 2020; López et al., 2018).

2.2. DCR graphs

DCR is a process modeling language belonging to the declarative paradigm (Hildebrandt & Mukkamala, 2011). The language is supported by a process modeling platform and commercial tools.[2] It comprises a core notation providing basic constructs for modeling DCR graphs. Fig. 1 illustrates a DCR graph describing the writing process of a project proposal. In this model, the activity "Download Submitted proposals" has no constraints and thus it can be executed several times and at any point in time. Similarly, the three weakly connected

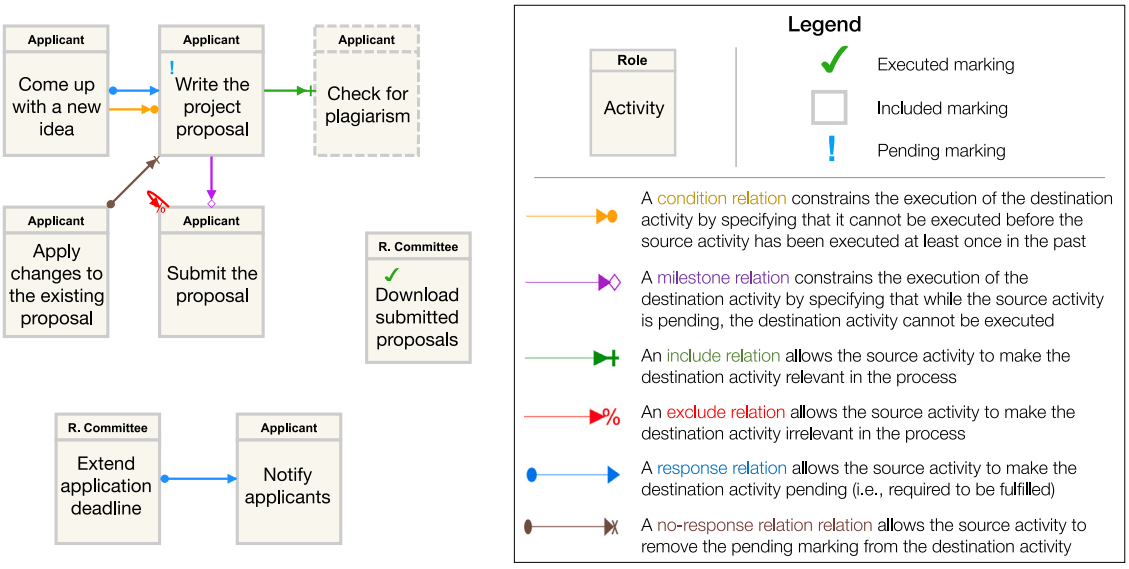


Fig. 1. A DCR graph describing the process of writing a project proposal.
Source: Adapted from Reichert and Weber (2012).

components of the model form independent blocks with no influence on each other.

Activities in DCR are characterized with markings recording their internal state. A marking comprises three Boolean values: *executed*, *included* and *pending*. The *executed* marking (indicated by a checkmark) indicates if an activity has been executed in the past (e.g., "Download submitted proposals"). The *included* marking (indicated by a solid border line) specifies whether an activity is relevant for the process at a particular point in the execution. The activities assigned this marking can constrain other activities (e.g., "Come up with a new idea"), while the activities missing this marking (depicted with a dashed border line) cannot (e.g., "Check for plagiarism"). Lastly, the *pending* marking (indicated by a blue exclamation mark) signifies that an activity must be executed before the process can end (e.g., "Write the project proposal").

As outlined in Section 2.1, the DCR language has 6 core relations. We use the terms "source activity" and "destination activity" to describe a pair of activities, connected with a DCR relation, depicted as a directed edge, running from the activity at the source of the edge to the activity at its destination. A *condition* relation (indicated by an orange edge) requires the source activity (e.g.,"Come up with a new idea") to be executed at least once before the destination activity (e.g., "Write the project proposal") can be executed. A *milestone* relation (indicated by a purple edge) specifies that as long as the source activity (e.g., "Write the project proposal") is required, the destination activity (e.g., "Submit the proposal") cannot be executed. The dynamic *inclusion* relation (indicated by a green edge) specifies that the source activity (e.g., "Write the project proposal") can activate the *included* marking of the destination activity (e.g., "Check for plagiarism"), making it *relevant* for the process. Conversely, the dynamic *exclusion* relation (indicated by a red edge) specifies that the source activity (e.g., "Submit the proposal") can disable the *included* marking of the destination activity (e.g., "Submit the proposal"), making it *irrelevant* for the process. The DCR graph shown in Fig. 1, links the activity "Submit the proposal" to itself using the exclusion relation, which, in turn, means that the activity can be executed only once. The *response* relation (indicated by a blue edge) denotes that the source activity (e.g., "Come up with a new idea") can activate the *pending* marking of the destination activity (e.g., "Write the project proposal"), making it *required* in the process. Lastly, the *no-response* relation (indicated by a brown edge) denotes that the source activity (e.g., "Apply changes to the existing proposal") can disable the *pending* marking of the destination activity (e.g., "Write the project proposal"), making it *no longer required* in the process (Andaloussi, Davis, Burattin, López, Slaats, & Weber, 2020). Note that it is possible to connect activities by more than one relation. In this case, the combined effect of the relations is applied. The process models used in the empirical study are based on this core notation (cf. Section 5).

2.3. Cognitive load theory

When humans are performing mental tasks, the demand imposed on their working memory is referred to in the literature as *Cognitive load* (Chen et al., 2016; Paas et al., 2003). According to the cognitive load theory (CLT), the limited capacity of the working memory acts as a bottleneck for humans when engaging with challenging tasks, thereby, affecting their performance negatively and leading them to error-prone situations (Chen et al., 2016; Paas et al., 2003). CLT discerns three types of cognitive load: *intrinsic*, *extraneous* and *germane* (Paas et al., 2003; Sweller, 2010). The *essential complexity* (Antinyan, 2020; Jr., 1987) inherent in the material being processed gives rise to *intrinsic load* (Paas et al., 2003; Sweller, 2010). In the context of process model comprehension, users would experience increased intrinsic load when the model encodes semantics describing the behavior of a complex business process. The *accidental complexity* (Antinyan, 2020; Jr., 1987) associated with the representation of the material being processed gives rise to *extraneous load* (Paas et al., 2003; Sweller, 2010). When engaging with process models having a poor visual representation, users are likely to experience this type of cognitive load. Lastly, the ability to construct mental schemes for an efficient processing of the material in hand denotes *germane load* (Paas et al., 2003; Sweller, 2010). The lack of abstraction skills and background in declarative process modeling could, for instance, induce higher germane load.

The metrics investigated in this study are likely to impact the intrinsic component of cognitive load as they capture structural aspects depending to a large extent on the complexity of the process behavior encoded in the model. Therefore, in our study, we shed light on this particular type of cognitive load. We do that through a controlled experiment where the factors affecting extraneous and germane loads are held as constant as possible in contrast to those

[1] A weakly connected component denotes a maximal sub-graph where the nodes are linked with some path regardless of the direction of the edges.

[2] See https://www.dcrgraphs.net and https://dcrsolutions.net.

susceptible to affect intrinsic load which are manipulated to study their impact (cf. Section 5).

2.4. Cognitive load measurements

Several measures have been proposed to estimate cognitive load in the literature (Chen et al., 2016). These measures have been organized into *subjective, performance, behavioral* and *physiological* (Chen et al., 2016). Subjective measures reflect humans' own perception of cognitive load (Chen et al., 2016; Sweller, Ayres, & Slava, 2011). Performance measures estimate the impact of cognitive load on the accomplishment of a given task (Chen et al., 2016). Behavior measures infer humans' cognitive load from their voluntary response to increased task difficulty (Chen et al., 2016). Lastly, physiological measures capture the changes in humans' physiological states due to variations of cognitive load (Kramer, 1991). In the following paragraphs, these classes of cognitive load measures are introduced together with the underlying theories, benefits, limitations, use in our empirical investigation and existing applications in model comprehension studies.

Subjective Measures. Cognitive load is associated with a subjective feeling of exertion that can be recognized by users and reported by the means of introspection (Chen et al., 2016). Questionnaires based on Likert scales are typically administered, in this vein, to inform about the difficulty and challenges perceived by users when engaging with a mental task (Sweller et al., 2011). Similarly, we use a self-assessment questionnaire of cognitive load in our study to capture users' *perceived difficulty*. In the literature, the measures extracted from self-assessment questionnaires were reported to be sensitive to different cognitive load levels (Sweller et al., 2011). However, these measures can be influenced by the social desirability bias and false metacognitive judgments (Cook, 2009).

Questionnaires based on the self-assessment of perceived difficulty have been widely used in the literature. In model comprehension studies, e.g., Figl and Laue (2011, 2015), perceived difficulty has been shown to vary depending on the task type (Figl & Laue, 2011), the control-flow patterns (Figl & Laue, 2015), the interactivity between the elements of the model (Figl & Laue, 2011) and the expertise of the modelers (Figl & Laue, 2015).

Performance Measures. Performance measures can serve as a proxy for cognitive load under the assumption that increased mental effort puts more strains on the working memory, which in turn, cause a drop in performance (Chen et al., 2016). Measures like *answer correctness* and *answering time* have been used in a way that reduced answer correctness and increased answering time reflect a decay in performance (Chen et al., 2016). Similarly, we use these two measures to estimate users' performance when solving comprehension tasks on process models.

Answer correctness and answering time are not intrusive since they do not require any additional input from the user when performing a task. However, their ability to discriminate different levels of cognitive load might be limited. Indeed, according to Veltman and Jansen (2005), users' performance decreases only in the states of cognitive *under*load (e.g., caused by boredom or lack of motivation) and cognitive *over*load (e.g., caused by very difficult tasks), while it remains constant in the states in between. This is because users can put more effort to compensate for increasing levels of cognitive load, provided that they do not experience cognitive overload.

Performance measures have been widely used in model comprehension studies, e.g., Bera (2012) and Sánchez-González et al. (2010). For instance, answer correctness was linked with model size (Sánchez-González et al., 2010). Likewise, answering time was associated with model density (Sánchez-González et al., 2010) and the quality of the visual notation (Bera, 2012).

Behavioral Measures. *Fixations* and *saccades* are voluntary eye-movements reflecting humans' cognitive processes (Holmqvist et al., 2011; Just & Carpenter, 1980; May, Kennedy, Williams, Dunlap, & Brannan, 1990; Meghanathan, van Leeuwen, & Nikolaev, 2015). A fixation denotes the timespan when the eye remains still at a position of the stimulus (Holmqvist et al., 2011). According to the eye-mind hypothesis (Just & Carpenter, 1980), in visual tasks, the eyes are fixating what the mind processes (Just & Carpenter, 1980). Consequently, longer and recurrent fixations can be associated with increased information processing. Among the common measures used to capture eye-fixation features in relation to cognitive load are *total fixation duration* (i.e., the total duration of all fixations in a specific time interval), *average fixation duration* (i.e., the average duration of all fixations in a specific time interval) and *fixation count* (i.e., the count of fixations in a specific time interval) (Just & Carpenter, 1980; Korbach, Brünken, & Park, 2017; Meghanathan et al., 2015). For all these measures, it is assumed that increased values indicate a high cognitive load.

Saccades are rapid eye movements between pairs of successive fixations (Holmqvist et al., 2011). This oculomotor event has been associated in the literature with cognitive load (Keskin, Ooms, Dogru, & De Maeyer, 2020; May et al., 1990). In particular, the *saccadic amplitude* measure, which denotes the distance traveled by the eye, has been shown to decrease as a response to increased task difficulty (Keskin et al., 2020; May et al., 1990). This phenomenon can be explained by the shrinkage of the human's visual field (or the so-called "tunnel vision" effect) to cope with memory overload. Indeed, when the amount of information exposed within a viewing area exceeds the typical processing capacity of an individual, the shrinkage of the visual field (manifesting in short saccades) helps to reduce the amount of information to be processed (Mackworth, 1965; May et al., 1990).

We rely on the aforementioned fixation and saccade-based measures because they are sensitive to different levels of cognitive load (May et al., 1990; Meghanathan et al., 2015) and are not intrusive since eye-trackers can be placed at a distance from the user. Nevertheless, it is important to ask the participants to follow a series of instructions (Holmqvist et al., 2011) to avoid confounding factors that can affect the reliability of these measures (e.g., miscalculation of gaze points due to head movements).

Fixation and saccade-based measures have been proposed to investigate users' mental effort during model comprehension tasks, e.g., (Petrusel, Mendling, & Reijers, 2016; Wang, Chen, Indulska, Sadiq, & Weber, 2022) and Zimoch et al. (2017). For instance, fixation duration was adopted in Wang et al. (2022) to compare users' mental effort when engaging with process models combined with linked or separated rules. Likewise, the counts of fixations and saccades were used in Zimoch et al. (2017) to investigate the difference between novice and intermediate modelers.

Physiological Measures. The human nervous system is composed of *central* and *peripheral* systems. The peripheral system has two branches: *somatic* and *autonomic* (Riedl & Léger, 2016). While the former is responsible for movements and for transmitting sensory information, the latter unconsciously regulates the body functions to cope with changes in mental or physical demands. The *autonomic* nervous system has two divisions: *sympathetic* and *parasympathetic*. The sympathetic division activates the human body when the level of mental or physical demand increases, whereas the parasympathetic one relaxes the body when the level of demand decreases. These events cause biological reactions like *pupil dilation* and changes in the *electrodermal activity* (*EDA*). Such reactions have been associated in the literature with an increase of cognitive load (Winter, Pryss, Probst, & Reichert, 2020).

Pupil size is captured by eye tracking devices and changes in pupil dilation can be derived through the *low/high index of pupillary activity* (Duchowski, Krejtz, Gehrer, Bafna, & Bækgaard, 2020). This measure differentiates the low and high frequencies of pupil oscillations to compute an index of cognitive load. The closer this index is to zero, the higher the cognitive load (Duchowski et al., 2020). As for EDA,

Table 1
Representative set of notable studies on process modeling guidelines.

Paradigm	Type	References
Imperative modeling	Studies	Becker, Rosemann, and Uthmann (2000), Corradini et al. (2018), Corradini, Polini, Re, Rossi, and Tiezzi (2022), Fischer (2010), Kopp (2022), Krogstie (2016), Mendling, Reijers, and van der Aalst (2010), Schrepfer (2010), White and Miers (2008)
	Literature reviews	Avila, dos Santos, Mendling, and Thom (2020), de Oca, Snoeck, Reijers, and Rodríguez-Morffi (2015), Dikici, Turetken, and Demirors (2018), Figl (2017), Moreno-Montes de Oca and Snoeck (2014)
Declarative modeling	Studies	Andaloussi et al. (2020), López-Acosta and Simon (2022)

it is captured by Galvanic Skin Response (GSR) devices measuring the conductance of the skin. Changes in conductance depend on the amount of sweat secreted by the skin sweat glands (Critchley, 2002). EDA can be devised into *tonic* and *phasic* components. The tonic component provides the Skin Conductance Level (SCL), a slowing varying signal that takes tens of seconds to minutes to change (Weber, Fischer, & Riedl, 2021). The phasic component, in turn, provides the Skin Conductance Response (SCR), a rapidly changing signal that takes one to five seconds to reflect a response to a stimulus (Winter et al., 2020). The SCR signal has several features, notably the number of peaks (i.e, SCR peaks count), which has been shown to increase with increasing mental demands (Winter et al., 2020).

Based on the literature (Duchowski et al., 2020; Winter et al., 2020), we use the low/high index of pupillary activity and the SCR peaks count. These two measures are sensitive to different levels of cognitive load (Duchowski et al., 2020; Winter et al., 2020) and can provide reliable insights if used carefully (Kramer, 1991). However, pupillary activity is not only sensitive to cognitive load but also to other factors such as illumination (Kramer, 1991), while EDA-based measures could be affected by user's physical activity (e.g., movements of the hand equipped with the GSR device) (Kramer, 1991). To avoid these confounding factors, it is, therefore, important to collect these measures in controlled environments where illumination is controlled and participants are well instructed about the use of the devices.

In model comprehension studies, the SCR peaks count has been shown to significantly change when engaging with process models having different levels of complexity (Winter et al., 2020). With regard to the low/high index of pupillary activity, to the best of our knowledge, this measure has not yet been used in the process model comprehension literature. Nevertheless, the findings of a series of experiments testing the metric in other mental tasks (e.g., counting, n-back, text-copy tasks) (Duchowski et al., 2020) confirm that the metric is a good indicator of cognitive load.

All in all, different measures have been proposed to estimate cognitive load. While the majority of model comprehension studies are restricted to a few ones (usually subjective or performance-driven Figl, 2017), our study uses a multi-modal approach supported by a larger set of measures. Such an approach has many advantages, first and foremost, the use of different modalities provides a multi-perspective assessment of cognitive load, through which one can overcome the potential limitations of the individual measures and thus provide a more comprehensive empirical account of users' cognitive load. Additionally, the use of behavioral and physiological measures brings more objectivity allowing to mitigate the social desirability bias and false metacognitive judgments associated with subjective measures (Cook, 2009). Last but not least, the continuous nature of behavioral and physiological measures allows for collecting real-time data and thus evaluating cognitive load at a high rate and sensitivity level (Chen et al., 2016).

3. Related work

This section presents the related work. Section 3.1 addresses the quality of process models. Section 3.2 provides an overview about the existing complexity metrics. Section 3.3 summarizes the understandability measures used in the existing empirical studies. Section 3.4, positions our contributions with respect to the related work and emphasizes how they extend to the state-of-the-art literature.

3.1. Quality of process models

There exists a wide stream of literature studying process model quality (Dikici et al., 2018; Krogstie, 2016). Model-based characteristics associated with size (Mendling, Reijers, & Cardoso, 2007), structure (Turetken, Rompen, Vanderfeesten, Dikici, & van Moll, 2016; Zugal, 2013), layout (Petrusel et al., 2016), hierarchy (Corradini et al., 2022) are among the pertinent proprieties affecting the understandability of process models. Accordingly, a number of quality guidelines have emerged (cf. overview of representative studies in Table 1). However, the largest proportion of these guidelines are tailored to imperative languages and typically assume a sequence-flow representation of process models. This assumption does not hold for declarative languages, which use constraints to model business processes. Hence, the applicability of the existing guidelines remains questionable for declarative languages for which the number of specific guidelines is rather limited (Andaloussi et al., 2020; López-Acosta & Simon, 2022). In Andaloussi et al. (2020), the authors suggest a set of modeling guidelines for DCR graphs, while in López-Acosta and Simon (2022), the authors propose a number of features to enhance the visual representation of DCR graphs.

3.2. Complexity metrics

The literature comprises a wide array of metrics designed to operationalize the existing guidelines and the underlying model proprieties (Moreno-Montes de Oca & Snoeck, 2014; Petrusel et al., 2016). These metrics originate from fields like graph theory, software engineering and information theory (cf. overview of representative studies in Table 2). From graph theory, for instance, connectivity, degree of vertices and density (Mendling, 2007) have been reformulated to measure the complexity of process models. Likewise, from software engineering, McCabe's Cyclomatic Complexity (McCabe, 1976), Lines of Code (LOC) and Halstead Complexity Metric (Halstead et al., 1977), previously used for source-code, have been adapted to fit process models (Cardoso et al., 2006). When it comes to information theory, based on Shannon entropy (Shannon, 1948), a metric capturing the heterogeneity (i.e., variability) of modeling constructs in process models has been proposed in Mendling (2007).

As reported in Table 2, most of the existing metrics capture the complexity of imperative process models expressed using languages such as Business Process Modeling Notation (BPMN) (Object Management Group (OMG), 2010) and Event-driven Process Chains (EPC) (Keller et al., 1992). The effects of the proprieties, operationalized using these metrics, on users' comprehension of process models have been demonstrated in a number of empirical studies, e.g., (Mendling & Strembeck, 2008; Reijers & Mendling, 2010) and Sánchez-González et al. (2010). Similar insights are, however, lacking for declarative process models, which besides a single study (Marin et al., 2015) investigating metrics for Case Management Modeling and Notation (CMMN) (Object Management Group OMG, 2014) models, less has been done to quantify their complexity.

3.3. Understandability measures for process models

The model proprieties and the metrics used to operationalize them have been tested in several studies to perceive their effect on the

Table 2
Representative set of notable studies on process model metrics.

Paradigm	Focus (type)	References
Imperative modeling	Graph theory (studies)	Cardoso et al. (2006), Latva-Koivisto (2001), Mendling (2007), Sa et al. (2012)
	Software engineering (studies)	Cardoso et al. (2006), Gruhn and Laue (2007), Mendling (2007), Reijers (2003), Reijers and Vanderfeesten (2004)
	Information theory (studies)	Cardoso et al. (2006), Cheng (2008), Mendling (2007)
	Multi (literature reviews)	Marin (2017), Moreno-Montes de Oca and Snoeck (2014), Polančič and Cegnar (2017), Zhou, Zhang, Chen, and Liu (2023)
Declarative modeling		Marin, Lotriet, and Van Der Poll (2015)

Table 3
Representative set of notable studies using understandability measures for process models. Abbreviations: Understd.: Understandability. Ans. Corr.: Answer correctness. Ans. T.: Answering time. Fix.: Fixations. Sacc.: Saccades, Pupil: Pupillary response. Per. Dif.:Perceived difficulty, EDA: Electrodermal activity.

Ref.	Investigated topics/proprieties	Understd Measure
Mendling and Strembeck (2008)	Modeling knowledge/experience, model structure, textual content, modeling purpose, visual layout	Ans. Corr.
Reijers and Mendling (2010)	Modeling experience & model structure	Ans. Corr.
Sánchez-González et al. (2010)	Model structure	Ans. Corr.
Winter, Pryss, Probst, Baß, and Reichert (2022)	Model structure	Ans. Corr.
Figl, Di Ciccio, and Reijers (2020)	Model Representation	Ans. Corr.
Winter, Neumann, Pryss, Probst, and Reichert (2023)	Visual layout, model structure	Ans. Corr.
Winter et al. (2021)	Model structure, textual content	Ans. Corr.
Sánchez-González et al. (2010)	Model structure	Ans. T.
Bera (2012)	Visual layout	Ans. T.
Winter et al. (2022)	Model structure	Ans. T.
Figl et al. (2020)	Model Representation	Ans. T.
Winter et al. (2023)	Visual layout, model structure	Ans. T.
Winter et al. (2021)	Model structure, textual content	Ans. T.
Figl and Laue (2011)	Model structure, task type	Per. Dif.
Figl and Laue (2015)	Model structure, task type, modeling knowledge/ experience	Per. Dif.
Winter et al. (2022)	Model structure	Fix.
Wang et al. (2022)	Rule integration	Fix. & Sacc.
Zimoch et al. (2017)	Modeling knowledge/experience, modeling language	Fix. & Sacc.
Petrusel et al. (2016)	Visual layout	Fix. & Sacc.
Petrusel and Mendling (2013)	Task-relevant regions in process models	Fix. & Sacc.
Bera, Soffer, and Parsons (2019)	Visual layout, modeling language, attention & visual associations on task-relevant regions in process models	Fix. & Sacc.
Winter et al. (2023)	Visual layout, model structure	Fix. & Sacc.
Weber, Neurauter, Pinggera, Zugal, Furtner, Martini, et al. (2015)	Task type	Pupil
Dobesova and Malcik (2015)	Task type	Pupil
Winter et al. (2020)	Model structure	EDA

comprehension of process models (Figl, 2017). In this vein, understandability was perceived as a theoretical construct and was operationalized using different measures (cf. overview of representative studies in Table 3). Based on Figl's literature review findings (details in supplementary material of Figl, 2017), the majority of the studies have deployed performance measures such as answer correctness, e.g., (Mendling & Strembeck, 2008; Reijers & Mendling, 2010) and Sánchez-González et al. (2010), answering time, e.g., Bera (2012), Sánchez-González et al. (2010) and perceived difficulty, e.g., Figl and Laue (2011, 2015). Besides, fewer studies have used behavioral measures based on fixations and saccades, e.g., Bera et al. (2019), Petrusel and Mendling (2013), Petrusel et al. (2016), Wang et al. (2022), Winter et al. (2023, 2022) and Zimoch et al. (2017). As for physiological measures, pupillary response analysis has been suggested in a handful of studies, e.g., (Dobesova & Malcik, 2015) and Weber et al. (2015). However, only one notable study reports practical insights about the use of pupillary data in experiments (Dobesova & Malcik, 2015). For the EDA analysis, to the best of our knowledge, the only study in this direction (Winter et al., 2020) investigates whether the comprehension of process models with complex structure leads to an increased EDA.

The aforementioned operationalizations of understandability overlap to a large extent with the measures used to operationalize cognitive load (cf. Section 2.4). In this work, we use this latter construct considering the rich body of interdisciplinary research spanning over several decades of theoretical and empirical studies (Sweller et al., 2011). Nevertheless, through our multi-modal approach, we cover the common understandability measures (i.e., answer correctness, time, perceived difficulty) in addition to several behavioral and physiological measures which have remained largely unexplored in the process modeling literature so far.

3.4. Own contributions

Our work contributes to the state-of-the-art literature from different angles. Firstly, we address the lack of metrics capturing the complexity of declarative process models (i.e., highlighted in Section 3.2) through a metric suite capturing different model proprieties. We also validate our suite of metrics empirically and thus extend the body of existing research with new insights on how different declarative model properties affect users' cognitive load. In turn, this empirical validation paves the road for the development of new declarative process modeling guidelines, enriching the limited set of existing guidelines (i.e., highlighted in Section 3.1) and contributing to the design of declarative models with enhanced quality. Besides, the multi-modal approach supporting our empirical validation benefits from the existing insights on the applicability of existing measures, used to investigate the understandability of process models. Moreover, we propose a more comprehensive combination of measures (i.e. compared to the literature in Section 3.3) providing a multi-perspective view of users' cognitive load when engaging with declarative process models.

4. Metrics for declarative process models

The literature comprises large collections of complexity metrics (cf. Section 3.2). However, the majority of these metrics were tailored for imperative process models (Cardoso et al., 2006; Gruhn & Laue, 2007; Latva-Koivisto, 2001; Mendling, 2007; Reijers, 2003; Reijers & Vanderfeesten, 2004; Sa et al., 2012), while only a few attempts have been made for their declarative counterparts (Marin et al., 2015). Still, the metrics proposed for imperative process models capture proprieties (e.g., size, cyclicity, concurrency, density, separability, connector heterogeneity), which could be relevant to measure for declarative process models as well. This in turn, raises questions on the applicability and adaptability of existing metrics to declarative process models.

The body of existing complexity metrics can be divided into two branches. The first branch focuses on the sequence-flow explicitly depicted in imperative process models (e.g., cyclicity, concurrency metrics Mendling, 2007). These metrics are not applicable to declarative models as they do not represent the process sequence-flow explicitly. The second part captures the graph structure proprieties of the model. Since both imperative and declarative models are based on graphs, graph-based metrics like size, density, separability and connector heterogeneity (Mendling, 2007) can in principle apply to both imperative and declarative models. However, only the size metric can be directly adopted, while for the other metrics, adaptations are needed to account for the differences between the two paradigms. Unlike imperative models, their declarative counterparts can be divided into weakly connected graph components. As mentioned in Section 2.1, these components denote blocks of activities that can be executed with no influence from the activities within the other components of the graph.

In the following, we address **RQ1** (cf. Section 1). Based on the metrics in the Mendling suite (Mendling, 2007), we formally define our new variants while taking into account the particularity of weakly connected graph components in declarative process models. Additionally, we provide an example to illustrate the computation of the new metrics.

In the subsequent definitions, for a Graph G, A_G denotes its set of nodes (i.e., activities), while C_G denotes its set of *constraints*. We use $|X|$ to refer to the cardinality of a set X.

Size Metric. This metric originates from LOC which was first introduced in software engineering and transferred to process modeling (Cardoso et al., 2006; Marin, 2017; Mendling, 2007). For imperative models, size denotes the number of nodes in the model (Mendling, 2007). Similarly, we define the size of a declarative process model represented as a graph G to be the sum of its activities and constraints:

$$Size(G) = |A_G| + |C_G|$$

Density Metric. This metric can be attributed to the graph theory, but it has been also used in the process modeling literature (Mendling, 2007). Notably, for imperative process models, density was operationalized as the number of arcs over the number of nodes in the model (Mendling, 2007). Given that a declarative process model is not necessarily fully connected but can rather span over multiple weakly connected graph components, we define density as the maximum number constraints over the number of activities in the weakly connected components of the graph (cf. Section 2.2). Let the set of weakly connected components of G: $\{c1, \ldots, c_n\}$ be $Comp(G)$ and the *activities* and *constraints* in the weakly connected component $c \in Comp(G)$ be A_c and C_c respectively, then Graph G has the following density:

$$Density(G) = \max_{c \in Comp(G)} \frac{|C_c|}{|A_c|}$$

Separability Metric. This metric relates to the concept of *cohesion* used in both software engineering and process modeling to capture the complexity of software artifacts (Mendling, 2007). Notably, for imperative process models, separability was defined as the number of cut vertices over the number of nodes in the model (Mendling, 2007). Since a declarative process model is not necessarily fully connected, we replace the notion of cut vertices with weakly connected components. Hence, we define the separability of a declarative process model, represented as a graph G, to be the number of weakly connected components over the number of activities and constraints in the model. We can formulate this definition as follows:

$$Separability(G) = \frac{|Comp(G)|}{|A_G| + |C_G|}$$

Constraint Variability Metric. This metric relates to Shannon entropy which captures uncertainty and randomness in the data (Shannon, 1948). In imperative process modeling, connector heterogeneity (Mendling, 2007) was defined as the entropy over the connector types incorporated in the model (Mendling, 2007). When it comes to declarative modeling, for more preciseness, we interchange the term "connector heterogeneity" with "constraint variability". In addition, we update the existing definition to incorporate the specificity of weakly connected components in declarative models. Hence, we define the constraint variability of a declarative process model represented as a graph G to be the maximum entropy over the different constraint types (cf. Section 2.1) in the components of the model.

Let $\mathcal{T}$ be the set of the different types of constraints in a declarative language (e.g., for a DCR graph, $\mathcal{T} = \{\mathtt{o}, \mathtt{m}, \mathtt{i}, \mathtt{e}, \mathtt{r}, \mathtt{n}\}$ with $\mathtt{o}$ for conditions, $\mathtt{m}$ for milestones, $\mathtt{i}$ for includes, $\mathtt{e}$ for excludes, $\mathtt{r}$ for responses, $\mathtt{n}$ for no-responses). Also, let $\mathcal{T}_c$ be the set containing the different types of constraints within a component c (i.e., in the graph G) and let C_c^t be the constraints of type t in component c. Then, the relative frequency p is:

$$p(c,t) = \begin{cases} \frac{|C_c^t|}{|C_c|} & \text{if } |C_c| > 0 \\ 0 & \text{otherwise} \end{cases}$$

The entropy, in turn, is defined as the negative sum over the number of constraint types of $p(c,t) \cdot \log_{|\mathcal{T}|}(p(c,t))$. Note that the base of the log function should match the number of constraints types in the language (similar to Mendling, 2007).

$$ConstraintVariability(G) = \max_{c \in \{c' | c' \in Comp(G) \wedge |C_{c'}| > 0\}} \left\{ - \sum_{t \in \mathcal{T}_c} p(c,t) \cdot \log_{|\mathcal{T}|}(p(c,t)) \right\}$$

Note that this formula considers only the components with at least one constraint.

Example. The declarative process model depicted as a DCR graph in Fig. 1 has the following complexity: $Size(G_1) = 15$, $Density(G_1) = 1.2$, $Separability(G_1) = 0.2$, $ConstraintVariability(G_1) = 1$.

5. Metrics validation: Empirical study

This section presents the empirical study designed to validate our metrics following the guidelines in Emp (2021). Section 5.1 presents the theoretical foundations based on which we formulate our hypotheses. Section 5.2 explains the study design. Section 5.3 summarizes the experiment procedure. Section 5.4 outlines the data processing and analysis approaches.

5.1. Hypotheses

Size. The number of elements in a declarative process model is likely to influence its understandability. This assumption is based on the cognitive load theory emphasizing the limited capacity of the human's working memory and thus the inability to cope with high information intake. Such intake is likely to affect humans' intrinsic load (Chen et al., 2016; Sweller et al., 2011). Accordingly, we formulate the following hypothesis: ***H1: Declarative process models with increased size yield higher intrinsic cognitive load than declarative process models with reduced size.***

Density. Increasing the ratio of constraints to activities in a model is likely to challenge its interpretation as this would raise the coupling between the model activities, requiring, in turn, extra checks to verify how each activity influences the rest of activities within the model. We conjecture that these checks would raise users' cognitive load considering the limited ability of declarative process models to offload computations compared to the imperative models. Indeed, while for imperative models the execution paths are explicitly depicted, when it comes to their declarative counterparts, the execution paths can be only inferred after evaluating the model constraints and the interplay between the process activities (Zugal, 2013). Accordingly, raising the ratio of constraints to activities would demand more mental computations and thus increased intrinsic cognitive load. Accordingly, we formulate the following hypothesis: ***H2. Declarative process models with increased density yield higher intrinsic cognitive load than declarative process models with reduced density.***

Separability. Declarative process models enable activities within a distinct weakly connected component to be executed independently, without affecting other components. In turn, verifying the impact of each activity on the rest of the model becomes simpler as its influence will be bounded to the component containing it. Accordingly, the more a model is partitioned into components, the easier it would be to comprehend. Conversely, if the same model has fewer components, it would be more cluttered and thus harder to interpret. This assumption finds support in the concept of *cognitive integration*, i.e., the process of combining information from different sources (Bera et al., 2019). Separability can influence the intrinsic cognitive load associated with this synthesis as the inability to isolate different parts of the model due to the presence of many dependencies between them would make the integration of information more difficult. Accordingly, we formulate the following hypothesis: ***H3. Declarative process models with reduced separability yield higher intrinsic cognitive load than declarative process models with increased separability.***

Constraint Variability. Using multiple constraint types in a declarative process model is likely to hinder its comprehension. Knowing that each type of constraints underlies a different semantics, users are required to remember and alternate between all these semantics before developing a good understanding of the model at hand. This task leads to the split attention effect where readers need to repeatedly move their attention between various concepts or pieces of information (Chandler & Sweller, 1992; Zugal, 2013). Such an effect can have a negative impact on users' intrinsic cognitive load. Accordingly, we formulate the following hypothesis: ***H4. Declarative process models with increased constraint variability yield higher intrinsic cognitive load than declarative process models with reduced constraint variability.***

Note that our hypotheses are framed within the context of comprehension tasks involving declarative process models in DCR.

5.2. Study design

This section provides an overview of the design of our study. Section 5.2.1 introduces our research model. Section 5.2.2 describes the used material. Section 5.2.3 provides an overview of the people who took part in the study.

5.2.1. Research model

An overview of our research model is provided in Fig. 2. The theoretical constructs in the treatment side are the proprieties measured by our *size, density, separability* and *constraint variability* metrics. Each construct (i.e., *factor*) comprises a "reduced level" and an "increased level". The former factor level is translated in models with low metric values, while the latter is translated in models with higher metric values. On the output side, the factor levels of the treatment are investigated with respect to their impact on cognitive load, which we operationalize using a variety of *subjective*, *performance*, *physiological* and *behavioral* measures as presented in Fig. 2.

We use a within-subject approach to design our experiment. Therein, each participant is repeatedly exposed to all factor levels. Such design supports repeated measurements as every participant provides data points for each factor level. Moreover, unlike the between-subject approach, the pairwise comparison of factor levels underlying our within-subject approach allows mitigating the risks associated with the heterogeneity of the participants and the individual differences expected to emerge as a result of experiencing different levels of germane cognitive load (cf. Section 2.3).

5.2.2. Material

The material used in the empirical study consists of a set of tasks. As illustrated in Fig. 3, each task contains a DCR graph, designed and verified for its syntax and semantic correctness within the DCR modeling platform.[3] Moreover, each task comprises a legend explaining the core DCR concepts, and an inference question. The complexity of the DCR graphs used in the different tasks reflects the reduced and increased levels of the factors presented in Section 5.2.1. To ensure that the effects we observe are caused by the elicited factor levels, a set of confounding factors are identified and addressed during the experiment's design phase. Based on the guidelines and modeling practices reported in the literature (Andaloussi et al., 2020; Zimoch, Pryss, Schobel & Reichert, 2017), we denote (a) *layout* as a confounding factor with a potential effect on extraneous load and (b) *modeling constructs*, (b) *relation patterns* and (c) *domain knowledge* as confounding factors with potential effects on intrinsic and germane loads.

We address the layout factor *(a)* by carefully setting up the visual representation of the models. Therein, the activities have the same dimensions, labeling style and spacing with the other activities and components of the model. As for edge crossing, 23 out of 24 of the used models do not have crossing edges, the only exception is one model where edge crossing was needed to account for the other layout parameters while increasing the density of the model. These parameters

[3] See https://www.dcrgraphs.net.

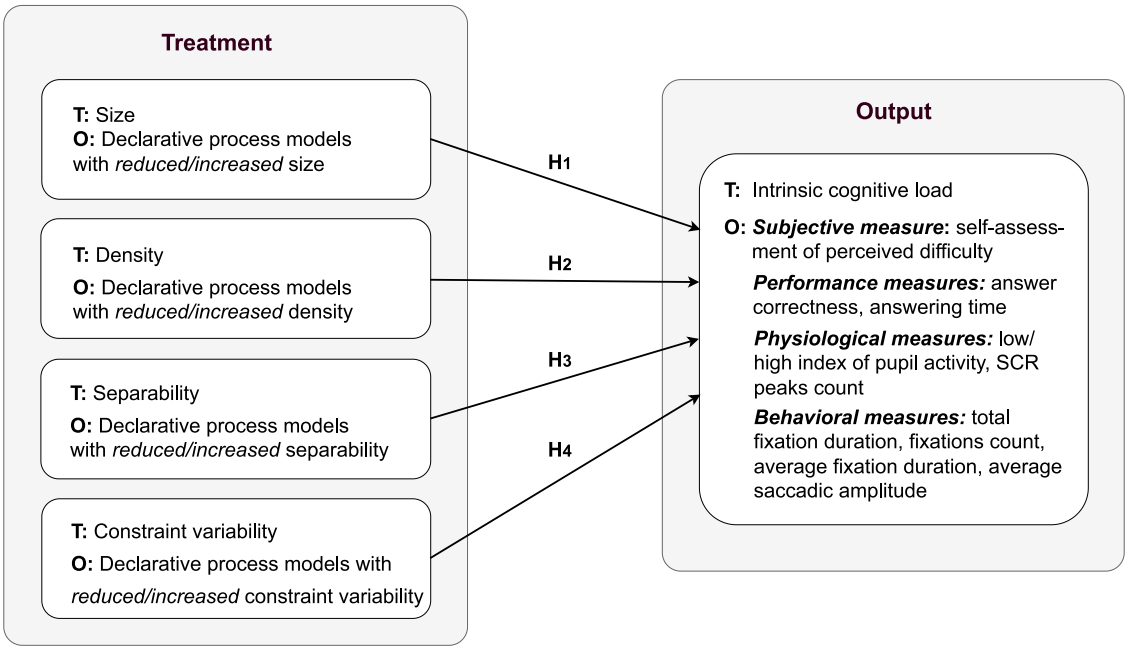


Fig. 2. Research model used in the empirical study: Abbreviations: T: Theoretical construct, O: Operationalization of construct.

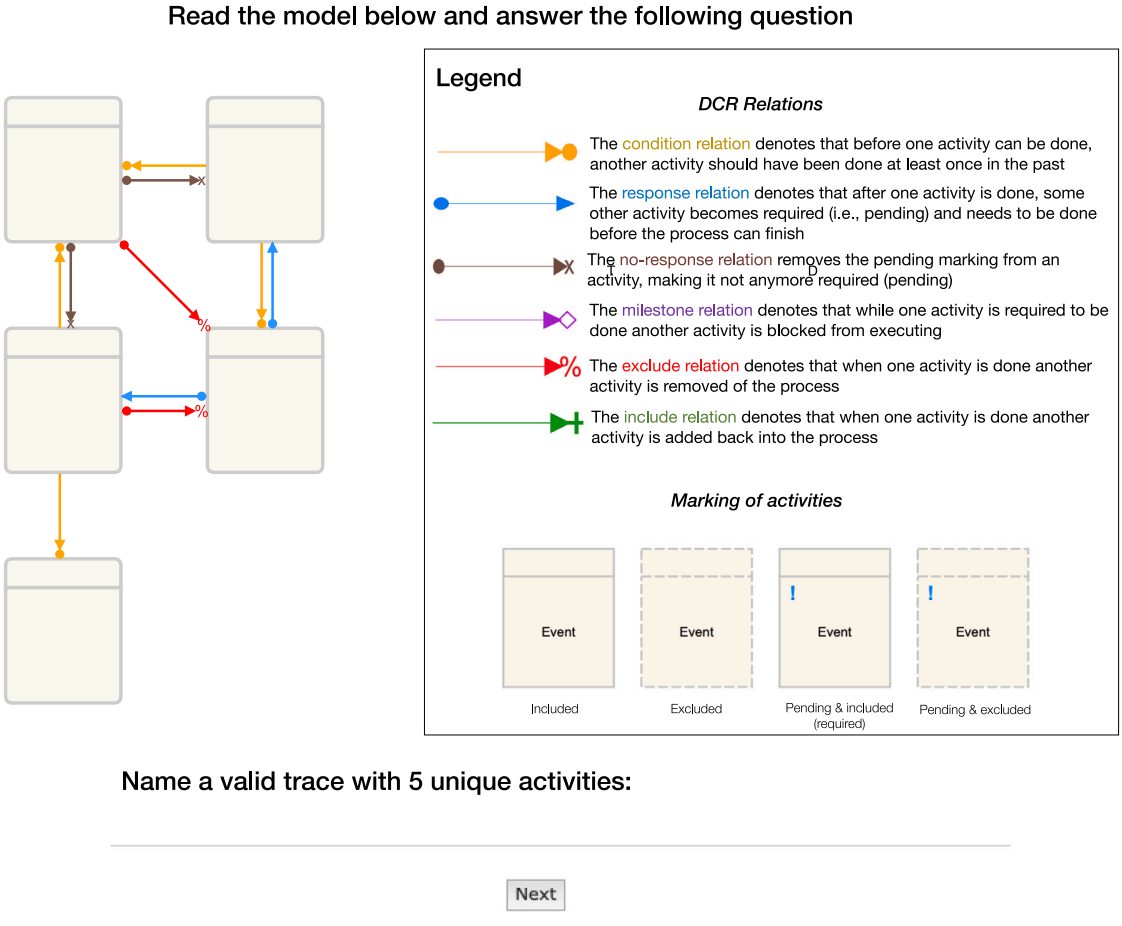


Fig. 3. An example of an experiment task. A higher resolution of this figure can be found at http://andaloussi.org/DeclarativeMetrics2023/Figures/taskPreview.pdf.

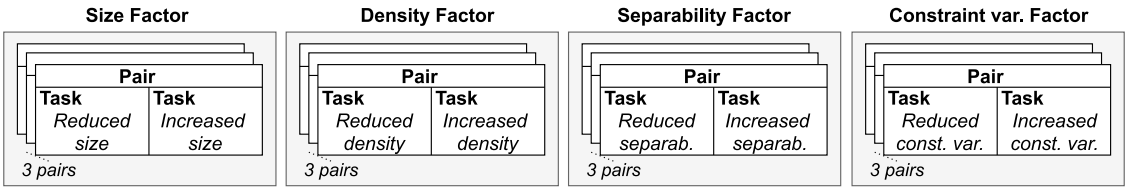


Fig. 4. Organization of Experiment Tasks. Abbreviations: Separab.: Separability, Const.: Constraint, Var.: Variability.

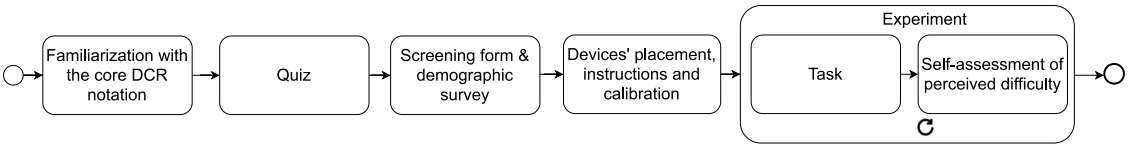


Fig. 5. Experiment procedure.

are detailed as part of our online appendix.[4] For the modeling constructs *(b)*, we limit the design of the models to the core DCR concepts while refraining from advanced notions like contextual process data, since they may have unwanted effects and thus confounding factors for the experiment. Regarding the relation patterns *(c)*, we rely solely on the 6 core DCR relations (cf. Section 2.2) which we use individually or in a combined manner to connect the models' activities. Lastly, we mitigate the influence of domain-knowledge *(d)* on participants' understanding of the process by anonymizing the model activities using random alphabet letters (similar to Pichler et al. (2011)).

The inference questions, in turn, ask the participants to provide a valid execution trace in which all the model activities are visited only once. Using these questions has several benefits. Indeed, it is more difficult to guess inference questions compared to the dichotomous ones (i.e., true or false questions). Moreover, since (1) our metrics measure the overall complexity of process models and (2) users tend to grant more attention to the parts relevant for answering the task compared to the irrelevant ones (Petrusel & Mendling, 2013), it is necessary to make the whole model relevant for the task. This approach ensures that the tasks used in the experiment capture the overall complexity of the model. Additionally, the used questions require checking the order in which activities are executed as well their dependencies and mutual influence. These aspects are crucial for understanding the control-flow underlying declarative process models (Zugal, 2013).

As shown in Fig. 4, 24 tasks are used to investigate our hypotheses. The tasks are organized into 4 sets focusing on different model proprieties (i.e., one of the following *factors*: size, density, separability, constraint variability). Each set contains 3 pairs, while each pair contains 2 tasks with models of which the complexity is the same — as measured by all the adapted metrics (cf. Section 4), except for the metric capturing the factor investigated within the set. Hence, within each pair, one task comprises a model with a reduced metric value for the investigated factor, while the other task has a model with an increased metric value. The grouping of tasks into pairs supports a pairwise comparison of the investigated factor levels as it will be explained in Section 5.4. The models together with the computed metrics and the used layout parameters are available online.[4] This material was thoroughly examined for any potential error before being used in the experiment.

5.2.3. Participants

Sixteen participants took part in the experiment. Their demographics are distributed as follows: 11 participants are male while 5 are female. 10 participants are within the age range of [20–30], 5 are within the range of [30–40] and one participant is above 40 years old. 13 participants have a background in Computer Science while the other 3 participants have backgrounds in Mathematics, Materials Engineering and Law respectively. With respect to their expertise in DCR graphs, on a 7-points scale asking to rate one's familiarity with DCR (i.e., 1 being unfamiliar with DCR to 7 being familiar with DCR), half of the participants fell within the range of [1–4], while others were in the range of [5–7].

5.3. Experiment procedure

The procedure used to run the experiment is illustrated in Fig. 5. Each session begins with a familiarization with DCR, followed by a quiz including 4 tasks on process models with different levels of complexity. The provided answers are discussed with the participant to validate their understanding of DCR and clarify any potential confusions. These steps are meant to test the participant's knowledge and ensure they have the minimum skills to take part in the experiment. Afterward, we administer a screening form and a demographic survey. The former form checks whether the participant has any vision issues or uses any equipment that can affect the collection of eye-tracking data (Holmqvist et al., 2011), while the latter, collects demographic information and ratings of the participant expertise in DCR graphs.

After filling out all the forms, an eye-tracking device (i.e., Tobii Pro X3–120[5]) mounted on a computer screen is placed in front of the participant. Then, a GSR device (i.e., Shimmer3 GSR[6]) is placed on the participant non-dominant hand. Following existing guidelines (Holmqvist et al., 2011; Imotions, 2017), the illumination in the lab is adjusted to mitigate any potential effect coming from the lighting condition in the room. Also, a calibration is conducted to ensure a correct mapping between gaze points and the stimulus. Additionally, the participant is asked to reduce head movements and keep the non-dominant hand still during the data collection.

The data collection is conducted in Tobii Pro Lab.[7] The tasks have no time restriction. To tackle the learning effect originating from the order in which questions are presented, the pairs and sets of tasks appear in random sequences to each participant. Following each task, a form with a Likert scale was administrated to collect the participant's self-assessment of perceived difficulty.

[4] See http://andaloussi.org/DeclarativeMetrics2023/.

[5] See https://www.tobiipro.com/product-listing/tobii-pro-x3-120/.

[6] See https://www.shimmersensing.com/products/shimmer3-wireless-gsr-sensor.

[7] See https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/Tobii-Pro-Lab-User-Manual/?v=1.145.

Table 4
Descriptive and inferential statistics.

H.	F.	Measure	Descriptive						Inferential
			Reduced l.			Increased l.			p-value
			N	Mean	SD	N	Mean	SD	
H1	Size	Perceived difficulty	48	1.854	0.85	48	4.271	1.047	**<.001**
		Answer correctness	48	1	0	48	0.833	0.377	**0.006**
		Answering time	48	30.91	15.637	48	153.349	68.073	**<.001**
		Low/high index of pupil activity	48	1.194	0.463	48	0.619	0.246	**<.001**
		SCR peaks count	42	3.048	2.409	42	12.143	9.421	**<.001**
		Total fixation duration	48	8250.854	6716.611	48	48462.583	34080.385	**<.001**
		Fixations count	48	51.833	32.771	48	309.479	183.115	**<.001**
		Average fixation duration	48	141.667	47.704	48	145.583	42.518	**0.008**
		Average saccadic amplitude	41	3.165	2.011	48	2.921	1.539	0.874
H2	Density	Perceived difficulty	48	2.396	0.765	48	3.979	1.021	**<.001**
		Answer correctness	48	0.792	0.41	48	0.75	0.438	0.283
		Answering time	48	56.583	26.326	48	122.702	85.22	**<.001**
		Low/high index of pupil activity	48	1.022	0.401	48	0.776	0.374	**<.001**
		SCR peaks count	42	5.762	6.525	42	9.333	9.573	**<.001**
		Total fixation duration	48	16405.75	11274.193	48	42196.313	35436.736	**<.001**
		Fixations count	48	107.479	61.081	48	269.229	226.095	**<.001**
		Average fixation duration	48	141.458	37.182	48	149.167	44.023	**<.001**
		Average saccadic amplitude	46	3.287	1.979	48	2.261	1.221	**<.001**
H3	Separability	Perceived difficulty	48	3.104	1.036	48	1.563	0.681	**<.001**
		Answer correctness	48	0.875	0.334	48	0.938	0.245	0.149
		Answering time	48	88.713	48.726	48	37.903	22.141	**<.001**
		Low/high index of pupil activity	48	0.811	0.372	48	1.176	0.413	**<.001**
		SCR peaks count	45	6.889	7.796	45	4.422	6.305	**<.001**
		Total fixation duration	48	28363.979	26016.1	48	9810.854	7542.482	**<.001**
		Fixations count	48	175.354	128.799	48	64.375	40.643	**<.001**
		Average fixation duration	48	144.708	41.697	48	140.333	39.601	**0.020**
		Average saccadic amplitude	47	2.595	1.327	43	3.737	1.913	**<.001**
H4	Const. variability	Perceived difficulty	48	1.917	0.846	48	3.188	0.891	**<.001**
		Answer correctness	48	0.938	0.245	48	0.708	0.459	**0.002**
		Answering time	48	52.109	28.598	48	84.663	42.157	**<.001**
		Low/high index of pupil activity	48	1.085	0.427	48	0.858	0.378	**0.002**
		SCR peaks count	45	4.378	5.019	45	7.133	9.034	**<.001**
		Total fixation duration	48	15205.125	12801.744	48	26676.938	21083.915	**<.001**
		Fixations count	48	98.792	69.757	48	169.25	111.037	**<.001**
		Average fixation duration	48	138.854	39.394	48	143.75	38.198	**0.008**
		Average saccadic amplitude	40	4.218	1.97	47	2.915	1.693	**<.001**

Notes: $p < 0.05$ informs that the pairwise difference of means between the two factor levels is significant. *Abbreviations:* Const.: Constraint. SD: Standard deviation. H.: Hypothesis, F.:Factor, l.: Level. *Means Units:* answering time: second, total fixations duration and average fixation duration: millisecond, average saccadic amplitude: visual degree.

5.4. Data processing and analysis

Data Processing. The data collected during the experiment were processed to extract the cognitive load measures shown in Fig. 2. These measures were computed at the level of each individual trial (i.e., a participant performing a task). Similar to Müller and Fritz (2016), the *self-assessment of perceived difficulty* was derived from participants' rantings' of perceived difficulty based on a 6-point Likert scale (with 1 being very easy to 6 being very difficult). The total number of participants resulted in 48 data points for each level of the investigated factors.

Answer correctness was defined as a binary score reflecting the correctness of the traces given by the participants. The total number of participants resulted in 48 data points for each level of the investigated factors.

Answering time was computed from the time interval separating the display of the task to the submission of the answer. The total number of participants resulted in 48 data points for each level of the investigated factors.

The low/high index of pupil activity was derived following the approach introduced in Duchowski et al. (2020). The approach, in a nutshell, computes the mean pupil diameter of the left and right eyes and relies on a wavelet decomposition to compute the ratio of low over high frequencies of pupil oscillations. The total number of participants resulted in 48 data points for each level of the investigated factors.

The *SCR peaks count* was computed using the SCR algorithm in Tobii Pro Lab,[7] which identifies and counts the number of peaks in the EDA signal. EDA was recorded from all the participants. However, a device failure occurred in a few trials. Hence, the total number of participants resulted in 42–45 data points for each level of the investigated factors (cf. Table 4).

The *fixation and saccade-based measures*, cover the *total fixations duration*, *fixations count*, *average fixation duration*, and *average saccadic amplitude*. These measures were computed using the "Tobii I-VT (fixation)" gaze filter implemented in Tobii Pro-lab.[7] The filter uses an eye-movement velocity threshold to discern fixations, saccades and their characteristics from the gaze data (Holmqvist et al., 2011). In our study, we derived the total fixations duration, fixations count and average fixation duration at the level of each individual trial (providing 48 data points per each measure/factor level), while we could not compute the average saccadic amplitude for all the trials as some of them did not comprise enough data points. The total number of participants resulted in 40–48 data points for each level of the investigated factors (cf. Table 4).

Data Analysis. During the analysis, descriptive statistics were computed for the reduced and increased levels of the size, density, separability and constraint variability factors. The results are presented on the left side of Table 4. Moreover, inferential tests were performed to validate our hypotheses. Therein, following a pairwise approach, the data points belonging to each factor level were compared across all

the used cognitive load measures (cf. Fig. 2). In total, we analyzed 36 paired data samples (4 factors × 9 measures) using the Wilcoxon signed-rank Test (single-tailed following the hypotheses formulation). We used this test for its general applicability to paired samples and its non-parametric nature obviating the requirement for normally distributed data. The results of the inferential tests are shown on the right side of Table 4.

6. Findings and discussion

This section summarizes the results of the descriptive and inferential statistics used to verify our hypotheses (cf. Section 6.1). In addition, it provides an overarching discussion comparing our findings with those in the literature and emphasizing their implication on future research, practice and educational settings (cf. Section 6.2).

6.1. Findings

The findings presented in this section address **RQ2** (cf. Section 1). The left side of Table 4 summarizes the descriptive statistics corresponding to the mean and standard deviation values of the reduced and increased factor levels investigated in the experiment (cf. Section 5.2.1). Overall the descriptive statistics show that when answering comprehension tasks incorporating models of *increased size*, *increased density*, *increased constraint variability* or *reduced separability*, the participants had higher perceived difficulty, answering time, SCR peaks count, total fixation duration, fixation count and average fixation duration. Moreover, they had lower answer correctness, low/high index of pupil activity and average saccadic amplitude. Based on the background provided in Section 2.4, the trends of all these measures suggest an increase in cognitive load.

The inferential statistics shown on the right side of Table 4 confirm this insight. Hypothesis **H1** (i.e., the influence of the size property on intrinsic cognitive load) is confirmed by all the measures except the average saccadic amplitude ($p = 0.874$). Hypothesis **H2** (i.e., the influence of the density property on intrinsic cognitive load) is confirmed by all the measures except the answer correctness ($p = 0.283$). Hypothesis **H3** (i.e., the influence of the separability property on intrinsic cognitive load) is confirmed by all the measures except the answer correctness ($p = 0.149$). Hypothesis **H4** (i.e., the influence of the constraint variability property on intrinsic cognitive load) is confirmed by all the measures.

6.2. Discussion

Overall, 33 out of 36 inferential tests support our hypotheses (cf. Table 4), which in turn show the effect of the model proprieties captured by our metrics on intrinsic cognitive load. Nevertheless, a few cognitive load measures could not support this conjecture. Namely, we could not show that model size has an effect on the participants' average saccadic amplitude. Also, we were unable to demonstrate the influence of density and separability on answer correctness. These findings are open to many interpretations. On the influence of model size on the average saccadic amplitude (i.e., part of **H1**), the models with increased size did not cause a shrink in the participants' visual field, which, in turn, manifested in large saccades with amplitudes similar to those occurring when engaging with models of reduced size. As mentioned in Section 2.4, our visual field shrinks when the information exposed within that field exceeds the limits of what we can process in real-time (Mackworth, 1965; May et al., 1990). This effect might not have happened because, at a local level, large models do not incorporate complex semantics (e.g., compared with dense models). Following this insight, one could speculate that when dealing with a large model, the local complexity of its semantics does not cause as much burden as the high amount of information that accumulates when trying to comprehend the model as a whole.

As for the influence of density on the answer correctness (i.e., part of **H2**), due to the absence of time restrictions when solving the given tasks, participants might have used enough time to cope with tasks involving dense models and thus provide correct answers regardless of their difficulty. This proposition finds evidence in the inferential tests reporting a significant difference in answering time when comparing tasks with reduced and increased densities (cf. Table 4). This difference is also visible in the descriptive statistics showing that participants spent twice as much time in models with increased density compared to those with reduced density.

The unclear influence of model separability on the answer correctness (i.e., part of **H3**) could be explained by the ceiling effect, occurring with high answering scores, leading to no significant difference between the investigated factor levels (Vogt & Johnson, 2011). This effect was also reported in a study investigating users' performance when engaging with process models (Zugal, 2013). In our context, the participants unexpectedly scored well on the tasks comprising complex process models having reduced separability (average answer correctness=0.875, cf. Table 4). This score is also the highest compared with the average answer correctness associated with the other complex models having increased size, density or constraint variability. Nevertheless, the other inferential tests in relation with **H3** confirm that reduced separability in declarative process models is still causing an increased difficulty, longer response time and is associated with several physiological and behavioral effects suggesting an increase of intrinsic cognitive load. This discrepancy could indicate that although the intrinsic cognitive load of the participants was high, it was still below the threshold of cognitive overload, which, in turn, did not affect the correctness of their answers. This proposition is supported by the authors in Veltman and Jansen (2005), who posit that despite an increase in cognitive load, users can maintain a constant performance as long as they put more effort to compensate for the high workload and do not attain cognitive overload.

The literature evaluates a wide array of metrics capturing model proprieties that are similar to those investigated in this work. With regard to the size metric, our findings line up with the results reported in Sánchez-González et al. (2010) where the size of BPMN models was associated with answer correctness and answering time. Conversely, our findings differ from those reported in Mendling and Strembeck (2008) as the authors could not correlate the size of EPC models and answer correctness. With regard to the density metric, our findings intersect with those reported in Sánchez-González et al. (2010) where the density of BPMN models was associated with answering time (Sánchez-González et al., 2010). Concerning the separability metric, our findings are similar to those reported in Figl and Laue (2011), Sánchez-González et al. (2010), where the effect of separability (e.g., in BPMN and EPC models) on answer correctness (Figl & Laue, 2011; Sánchez-González et al., 2010) could not be shown. This effect was nonetheless found in other studies (Mendling & Strembeck, 2008). For the constraint variability metric, similar to our findings, the authors in Sánchez-González et al. (2010) associated the underlying model property with answering time and answer correctness on BPMN models. Conversely, the authors in Mendling and Strembeck (2008), Reijers and Mendling (2010) could not link that property with answer correctness on EPC models.

The disparities between our results and the literature findings could be attributed to different factors, in particular, the difference of languages (e.g., BPMN, EPC, DCR), language paradigm (i.e., declarative, imperative) and the operationalization of the metrics. The design of the experiments and the analysis approaches could be other reasons for these differences. While the majority of the papers test the metrics altogether and use correlation analyses between the investigated complexity metrics and measures such as answer correctness, answering time and perceived difficulty, our study treats the model proprieties captured by each metric separately. Lastly, the used material could be

another cause for such differences not only between our study and the others but also among the other studies themselves.

Implications. The outcome of this work has implications on future research, practice and educational settings. Regarding research, our metrics serve to assess the quality of declarative process models and can be used to enhance existing declarative process mining algorithms (Slaats, 2020) as well as provide heuristics to improve the automatic discovery of low-complexity process models from event logs. Moreover, considering our multi-modal approach and the continuous nature of the measures used in the empirical study, the ability to consistently monitor cognitive load from different angles opens up for a new class of experiments with the potential to pinpoint both where and when cognitive load occurs. This is because fixations have spatial coordinates which can be linked to specific constructs of the process model. The same fixations can be investigated from a temporal perspective and thus be also linked with physiological events related to pupil dilation and EDA (e.g., operationalized through low/high index of pupillary activity and SCR peaks count). Hence, it might be possible to isolate any single construct of the process model and study the associated cognitive load. A handful of attempts have been reported in this direction within the software engineering field (Fakhoury, Roy, Ma, Arnaoudova, & Adesope, 2020; Hijazi, Couceiro, Castelhano, De Carvalho, Castelo-Branco, & Madeira, 2021), which in turn, demonstrate the viability of this approach and motivates its adaptation in the process modeling field for a more fine-grained analysis of modelers cognitive load.

With regard to practice, tool vendors can implement our metrics to deliver new intelligent modeling platforms, automating the evaluation of declarative process models at design time. Such an evaluation can tell how complex is a process model and provide indications about the workload and expertise levels required to maintain it. Additionally, given the continuous nature of the used measurements, with further development, one can investigate their ability to be used in online settings in the form of a tool-support that can detect and enact to high cognitive load during process modeling or comprehension tasks.

Education is also an important direction where our work could have implications. Similar to the existing imperative process modeling guidelines (e.g., Mendling et al., 2010), a new compendium of declarative process modeling guidelines can be derived based on our empirical findings. At a high level, modelers should always aim at reducing the size, density and constraint variability of declarative models, while increasing their separability. At an operational level, several recommendations can be formulated. For instance, modelers should keep in mind, the key principle behind declarative modeling that is to avoid a sequence-flow design and rather focus on the constraints guiding the overall process behavior. Besides supporting more flexibility, this guideline will result in models with fewer constraints, which in turn would reduce their density. Moreover, modelers should use a proper decomposition approach to divide the process specifications into distinct components or totally separate models. The former advice would increase the separability of the model, while the latter would reduce its size. As for our multi-modal approach, with further development, there might be room for transferring its research and practical implications to educational settings by developing novel tools supporting e-learning through cognitive load measures that can pinpoint where and when learners are challenged. This in turn, can help instructors to adjust the learning pace and the material.

7. Threats to validity

The following paragraphs discuss the aspects threatening the validity of our empirical study. We organize these aspects following the classification proposed in Wohlin, Runeson, Höst, Ohlsson, Regnell, and Wesslén (2012).

Internal Validity. The causality between the treatment and output variables can be threatened when the data is collected in uncontrolled environments or no clear instructions are provided to the participants. In our empirical study, we mitigated this threat by controlling the factors susceptible to bias the results of our experiment and instructing the participants following a clear and uniform protocol. The study design (i.e., models, tasks, participants) could also underlie some threats to internal validity. The DCR models used in the experiment were designed following the metrics defined in Section 4 to accurately capture the associated model properties. The concrete metric values for each of these models can be found in the appendix referred to in Section 5.2.2. Moreover, a set of model-related confounding factors were defined and addressed to further support the internal validity of our experiment (cf. Section 5.2.2). Nonetheless, in 1 out of the 24 used models, we experienced edge crossing as a result of increased density. To alleviate potential biases in the selection of process models, we have used three model replicates for each factor level to ensure that the observed effects are consistent across different models. Since the DCR models were designed and verified using the DCR modeling platform, we can ensure that they have no syntax or semantic errors. with regards to the used tasks, the associated comprehension questions were formulated with the same words to mitigate any task-related confounding factor. As for the participants, we provided a uniform familiarization ensuring that all participants (regardless of their expertise) have the minimum set of skills required to take part in the data collection. Moreover, we adopted a within-subject design to further mitigate the effect of participants' expertise on the obtained results. Furthermore, we randomized the sequences in which the tasks were displayed to the participants to mitigate the learning effect.

External Validity. Since the process models used in the empirical study were represented as DCR graphs, we may not have strong evidence to generalize our findings to other languages. However, it is clear that DCR has several synergies with other declarative languages such as Declare (Pesic et al., 2007) and thus it is plausible that our findings would apply to them too. Moreover, as we have focused only on the core notation of DCR, there might be limitations when trying to generalize our findings to models with advanced concepts. We focused on the core notation of DCR to avoid any potential confounding factors emerging from the use of other advanced concepts. Furthermore, the anonymization of activities using random letters of the alphabet may not reflect the way process models are represented in the real-world. However, we justify our design decision by the need to mitigate the effect of domain knowledge which we could have faced if we used domain-dependent models. When it comes to the design of our tasks, we acknowledge that our question type is not unique as users might be asked to extract different information from process models during comprehension tasks. However, the used question type require users to perform a set of checks to understand the order, the dependencies and the influence of the model activities on each other. These checks remain crucial when trying to interpret the control-flow of most process models (Zugal, 2013). Last but not the least, since we have collected the data in a laboratory environment, there is still need to verify our findings in industrial settings.

Construct Validity. The mono-method bias can threaten the generalization of the findings to the theories and concepts of the literature (Wohlin et al., 2012). To tackle this limitation, we used a multi-modal approach ensuring that our hypotheses can be verified using different measures.

Conclusion Validity. A threat to validity in this vein could be associated with our sample size. Nonetheless, our range is common in empirical studies deploying physiological or behavioral measures (Gulden, Burattin, Andaloussi, & Weber, 2019; Winter et al., 2020). Moreover, we have used a within-subject design with repeated measurements, allowing us to obtain from 40 to 48 data points per factor level (cf. Table 4). Furthermore, we may not be able to provide general statements about how complex is a declarative process model because the proposed metrics capture only a part of the possible complexity aspects.

8. Conclusion and future work

Size, density, separability and constraint variability are model proprieties that we have operationalized in this work into complexity metrics for declarative process models and investigated empirically in a controlled experiment. The findings demonstrate the impact of these model proprieties on cognitive load. Overall, this work advances the state-of-the-art literature, being the first to propose empirically validated metrics to measure the complexity of process models in the declarative paradigm and the first to use a comprehensive multi-modal approach to measure cognitive load in the process modeling literature.

There are several directions for future work. Tool vendors can implement our metrics in their modeling tools. However, it is important to consider a number of challenges and limitations in that regard and investigate how they can be mitigated. Notably, from an implementation point of view, although our complexity metrics are formally defined (cf. Section 4), integrating them into existing modeling tools might not be trivial as each tool has its own Application Programming Interfaces (i.e., APIs) and plugin architecture. From a user point of view, while our complexity metrics capture relevant properties of process models, as mentioned in Section 7, there is no guarantee that these metrics capture all possible angles of process complexity. Hence, users should not rely solely on them nor aim at over-optimizing them at the cost of other potentially important quality aspects.

Besides, large-scale studies (e.g., in classrooms, companies) involving comprehension, modeling and maintenance tasks on declarative process models other than DCR graphs can be performed to further investigate our metrics. Last but not least, the proposed multi-modal approach can be further exploited to provide a more fine-grained assessment of cognitive load showing exactly what constructs and modeling patterns are challenging the understanding of declarative models.

CRediT authorship contribution statement

Amine Abbad-Andaloussi: Conceptualization, Methodology, Investigation, Formal analysis, Validation, Writing – original draft. **Andrea Burattin:** Conceptualization, Methodology, Validation, Writing – review & editing. **Tijs Slaats:** Conceptualization, Methodology, Validation, Writing – review & editing. **Ekkart Kindler:** Conceptualization, Methodology, Validation, Writing – review & editing. **Barbara Weber:** Conceptualization, Methodology, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

An online appendix is provided in the paper

Acknowledgments

This work is supported by Innovation Fund Denmark project EcoKnow (Number: 7050–00034A) and by the International Postdoctoral Fellowship (IPF) from the University of St. Gallen, Switzerland (Number: 1031574). The funding organizations had no involvement in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to publish the article.

References

Andaloussi, A. A., Davis, C. J., Burattin, A., López, H. A., Slaats, T., & Weber, B. (2020). Understanding quality in declarative process modeling through the mental models of experts. In *International conference on business process management* (pp. 417–434). Springer.

Antinyan, V. (2020). Evaluating essential and accidental code complexity triggers by practitioners' perception. *IEEE Software*, *37*(6), 86–93.

Avila, D. T., dos Santos, R. I., Mendling, J., & Thom, L. H. (2020). A systematic literature review of process modeling guidelines and their empirical support. *Business Process Management Journal*.

Becker, J., Rosemann, M., & Uthmann, C. v. (2000). Guidelines of business process modeling. In *Business process management* (pp. 30–49). Springer.

Bera, P. (2012). Does cognitive overload matter in understanding BPMN models? *Journal of Computer Information Systems*, *52*(4), 59–69.

Bera, P., Soffer, P., & Parsons, J. (2019). Using eye tracking to expose cognitive processes in understanding conceptual models. *MIS Quarterly*, *43*(4), 1105–1126.

Cardoso, J., Mendling, J., Neumann, G., & Reijers, H. A. (2006). A discourse on complexity of process models. In *International conference on business process management* (pp. 117–128). Springer.

Chandler, P., & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology*, *62*(2), 233–246.

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., et al. (2016). *Robust multimodal cognitive load measurement*. Springer.

Cheng, C.-Y. (2008). *Complexity and usability models for business process analysis*. The Pennsylvania State University.

Cook, A. E. (2009). Measurement of cognitive load during multimedia learning activities. In *Cognitive effects of multimedia learning* (pp. 34–50). IGI Global.

Corradini, F., Ferrari, A., Fornari, F., Gnesi, S., Polini, A., Re, B., et al. (2018). A guidelines framework for understandable BPMN models. *Data & Know Engineering*, *113*, 129–154.

Corradini, F., Polini, A., Re, B., Rossi, L., & Tiezzi, F. (2022). Consistent modelling of hierarchical BPMN collaborations. *Business Process Management Journal*, *28*(2), 442–460.

Critchley, H. D. (2002). Electrodermal responses: what happens in the brain. *The Neuroscientist*, *8*(2), 132–142.

de Oca, I. M.-M., Snoeck, M., Reijers, H. A., & Rodríguez-Morffi, A. (2015). A systematic literature review of studies on business process modeling quality. *Information and Software Technology*, *58*, 187–205.

Dikici, A., Turetken, O., & Demirors, O. (2018). Factors influencing the understandability of process models: A systematic literature review. *Information and Software Technology*, *93*, 112–129.

Dobesova, Z., & Malcik, M. (2015). Workflow diagrams and pupil dilatation in eye-tracking testing. In *ICETA' 2015* (pp. 1–6). IEEE.

Duchowski, A. T., Krejtz, K., Gehrer, N. A., Bafna, T., & Bækgaard, P. (2020). The low/high index of pupillary activity. In *CHI conference on human factors in computing systems* (pp. 1–12).

Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2013). *Business process management*.

Emperical Standard Guidelines for Experiments (2021), URL https://github.com/acmsigsoft/EmpiricalStandards/blob/master/docs/Experiments.md.

Fahland, D., Lübke, D., Mendling, J., Reijers, H., Weber, B., Weidlich, M., et al. (2009). Declarative versus imperative process modeling languages: The issue of understandability. In *Proceedings of EMMSAD* (pp. 353–366).

Fakhoury, S., Roy, D., Ma, Y., Arnaoudova, V., & Adesope, O. (2020). Measuring the impact of lexical and structural inconsistencies on developers' cognitive load during bug localization. *Empirical Software Engineering*, *25*(3), 2140–2178.

Figl, K. (2017). Comprehension of procedural visual business process models. *Business & Information Systems Engineering*, *59*(1), 41–67.

Figl, K., Di Ciccio, C., & Reijers, H. A. (2020). Do declarative process models help to reduce cognitive biases related to business rules? In *Conceptual modeling: 39th international conference, ER 2020, Vienna, Austria, November 3–6, 2020, proceedings 39* (pp. 119–133). Springer.

Figl, K., & Laue, R. (2011). Cognitive complexity in business process modeling. In *International conference on advanced information systems engineering* (pp. 452–466). Springer.

Figl, K., & Laue, R. (2015). Influence factors for local comprehensibility of process models. *International Journal of Human-Computer Studies*, *82*, 96–110.

Fischer, L. (2010). *Vol. 1, BPMN 2.0 handbook first edition: foreword by bruce silver*. Future Strategies Inc..

Gruhn, V., & Laue, R. (2007). Approaches for business process model complexity metrics. In *Technologies for business information systems* (pp. 13–24). Springer.

Gulden, J., Burattin, A., Andaloussi, A. A., & Weber, B. (2019). From analytical purposes to data visualizations: a decision process guided by a conceptual framework and eye tracking. *Software & Systems Modeling*, 1–24.

Halstead, M. H., et al. (1977). *Vol. 7, Elements of software science*. Elsevier New York.

Hijazi, H., Couceiro, R., Castelhano, J., De Carvalho, P., Castelo-Branco, M., & Madeira, H. (2021). Intelligent biofeedback augmented content comprehension (TellBack). *IEEE Access*, *9*, 28393–28406.

Hildebrandt, T. T., Andaloussi, A. A., Christensen, L. R., Debois, S., Healy, N. P., López, H. A., et al. (2020). EcoKnow: Engineering effective, co-created and compliant adaptive case management systems for knowledge workers. In *Proceedings of the international conference on software and system processes* (pp. 155–164).

Hildebrandt, T. T., & Mukkamala, R. R. (2011). Declarative event-based workflow as distributed dynamic condition response graphs. *EPTCS*, *69*, 59–73, arXiv:1110.4161.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: a comprehensive guide to methods and measures*. OUP Oxford.

Imotions (2017). Galvanic skin response: The complete pocket guide.

Jr., F. P. B. (1987). No silver bullet - essence and accidents of software engineering. *Computer*, *20*(4), 10–19. http://dx.doi.org/10.1109/MC.1987.1663532.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329.

Keller, G., Nüttgens, M., & Scheer, A. W. (1992). *Semantische prozessmodellierung auf der grundlage "ereignisgesteuerter prozessketten (EPK)"*.

Keskin, M., Ooms, K., Dogru, A. O., & De Maeyer, P. (2020). Exploring the cognitive load of expert and novice map users using eeg and eye tracking. *ISPRS International Journal of Geo-Information*, *9*(7).

Kopp, A. (2022). Guidelines and a software tool for quality assessment of BPMN business process models. *Journal of Emerging Technologies*, *2*(2), 55–65.

Korbach, A., Brünken, R., & Park, B. (2017). Measurement of cognitive load in multimedia learning: a comparison of different objective measures. *Instructional Science*, *45*(4), 515–536.

Kramer, A. F. (1991). Physiological metrics of mental workload. *Multiple Task Performance*, 279–328.

Krogstie, J. (2016). *Quality in business process modeling*. Cham: Springer.

Latva-Koivisto, A. M. (2001). Finding a complexity measure for business process models. *Technical report*, Helsinki University of Technology.

López, H. A., Debois, S., Hildebrandt, T. T., & Marquard, M. (2018). The process highlighter: From texts to declarative processes and back. *2196*, In *CEUR workshop proceedings, BPM 2018 demo* (pp. 66–70).

López-Acosta, H.-A., & Simon, V. D. (2022). How to (re) design declarative process notations? A view from the lens of cognitive effectiveness frameworks. In *15th IFIP working conference on the practice of enterprise modeling 2022*. CEUR-WS.

Mackworth, N. H. (1965). Visual noise causes tunnel vision. *Psychonomic Science*, *3*(1), 67–68.

Marin, M. A. (2017). *Exploring complexity metrics for artifact-centric business process models* (Ph.D. thesis), University of South Africa (South Africa).

Marin, M. A., Lotriet, H., & Van Der Poll, J. A. (2015). Metrics for the case management modeling and notation (CMMN) specification. In *Proceedings of the 2015 annual research conference on South African institute of computer scientists and information technologists* (pp. 1–10).

Marquard, M., Shahzad, M., & Slaats, T. (2016). Web-based modelling and collaborative simulation of declarative processes. In *Business process management* (pp. 209–225). Cham: Springer International Publishing.

May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye movement indices of mental workload. *Acta Psychologica*, *75*(1), 75–89.

McCabe, T. J. (1976). A complexity measure. *IEEE Transactions on Software Engineering*, (4), 308–320.

Meghanathan, R. N., van Leeuwen, C., & Nikolaev, A. R. (2015). Fixation duration surpasses pupil size as a measure of memory load in free viewing. *Frontiers in Human Neuroscience*, *8*, 1063.

Mendling, J. (2007). *Detection and prediction of errors in EPC business process models* (Ph.D. thesis), Wirtschaftsuniversität Wien Vienna.

Mendling, J., Reijers, H. A., & van der Aalst, W. M. (2010). Seven process modeling guidelines (7PMG). *Information and Software Technology*, *52*(2), 127–136.

Mendling, J., Reijers, H. A., & Cardoso, J. (2007). What makes process models understandable? In *International conference on business process management 2007* (pp. 48–63).

Mendling, J., & Strembeck, M. (2008). Influence factors of understanding business process models. In *International conference on business information systems* (pp. 142–153). Springer.

Müller, S. C., & Fritz, T. (2016). Using (bio) metrics to predict code quality online. In *2016 IEEE/ACM 38th international conference on software engineering* (pp. 452–463). IEEE.

Object Management Group (OMG) (2010). Business process modeling notation 2.0. URL https://www.omg.org/spec/BPMN/2.0/.

Object Management Group OMG (2014). Case Management Model and Notation. URL https://www.omg.org/spec/CMMN/.

Moreno-Montes de Oca, I., & Snoeck, M. (2014). Pragmatic guidelines for business process modeling. Available at SSRN 2592983.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71.

Pesic, M., Schonenberg, H., & van der Aalst, W. M. (2007). DECLARE: Full support for loosely-structured processes. In *EDOC' 2011* (p. 287).

Petri, C. A. (1962). Kommunikation mit Automaten. *Fakultät FÜR Mathematik Und Physik*, *Doktor*, 128.

Petrusel, R., & Mendling, J. (2013). Eye-tracking the factors of process model comprehension tasks. In *International conference on advanced information systems engineering* (pp. 224–239). Springer.

Petrusel, R., Mendling, J., & Reijers, H. A. (2016). Task-specific visual cues for improving process model understanding. *Information and Software Technology*, *79*, 63–78.

Pichler, P., Weber, B., Zugal, S., Pinggera, J., Mendling, J., & Reijers, H. A. (2011). Imperative versus declarative process modeling languages: An empirical investigation. In *BPM' 2011*. Springer.

Polančič, G., & Cegnar, B. (2017). Complexity metrics for process models–a systematic literature review. *Computer Standards & Interfaces*, *51*, 104–117.

Reichert, M., & Weber, B. (2012). *Enabling flexibility in process-aware information systems*. Springer.

Reijers, H. (2003). A cohesion metric for the definition of activities in a workflow process. *Vol. 1*, In *Proceedings of the eighth CAiSE/IFIP8* (pp. 116–125).

Reijers, H. A., & Mendling, J. (2010). A study into the factors that influence the understandability of business process models. *IEEE Transactions on Systems, Man, and Cybernetics*, *41*(3), 449–462.

Reijers, H. A., & Vanderfeesten, I. T. (2004). Cohesion and coupling metrics for workflow process design. In *International conference on business process management* (pp. 290–305). Springer.

Riedl, R., & Léger, P.-M. (2016). Fundamentals of neuroIS. *Studies in Neuroscience, Psychology and Behavioral Economics*, 127.

Sa, L., Garci, F., Ruiz, F., & Mendling, J. (2012). A study of the effectiveness of two threshold definition techniques. In *IET conference proceedings*. The Institution of Engineering & Technology.

Sánchez-González, L., García, F., Mendling, J., & Ruiz, F. (2010). Quality assessment of business process models based on thresholds. In *OTM' 2010* (pp. 78–95). Springer.

Schrepfer, M. (2010). Modeling guidelines for business process models. *Humboldt-Universitat Zu Berlin*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Slaats, T. (2020). Declarative and hybrid process discovery: Recent advances and open challenges. *Journal on Data Semantics*, *9*(1).

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*(2), 123–138.

Sweller, J., Ayres, P., & Slava, K. (2011). *Cognitive load theory*. Springer.

Turetken, O., Rompen, T., Vanderfeesten, I., Dikici, A., & van Moll, J. (2016). The effect of modularity representation and presentation medium on the understandability of business process models in BPMN. In *International conference on business process management* (pp. 289–307). Springer.

Veltman, J., & Jansen, C. (2005). *The role of operator state assessment in adaptive automation*. Tno Defence Security and Safety Soesterberg (Netherlands).

Vogt, W. P., & Johnson, B. (2011). *Dictionary of statistics & methodology*. Sage.

Wang, W., Chen, T., Indulska, M., Sadiq, S., & Weber, B. (2022). Business process and rule integration approaches—An empirical analysis of model understanding. *Information Systems*, *104*, Article 101901.

Weber, B., Fischer, T., & Riedl, R. (2021). Brain and autonomic nervous system activity measurement in software engineering: A systematic literature review. *Journal of Systems and Software*, *178*.

Weber, B., Neurauter, M., Pinggera, J., Zugal, S., Furtner, M., Martini, M., et al. (2015). Measuring cognitive load during process model creation. In *Information systems and neuroscience*.

White, S. A., & Miers, D. (2008). *BPMN modeling and reference guide: understanding and using BPMN*. Future Strategies Inc..

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 Systems with Applications*, *213*, Article 119217.

Winter, M., Pryss, R., Probst, T., Baß, J., & Reichert, M. (2022). Measuring the cognitive complexity in the comprehension of modular process models. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(1), 164–180.

Winter, M., Pryss, R., Probst, T., & Reichert, M. (2020). Towards the applicability of measuring the electrodermal activity in the context of process model comprehension: Feasibility study. *Sensors*, *20*(16).

Winter, M., Pryss, R., Probst, T., Schlee, W., Tallon, M., Frick, U., et al. (2021). Are non-experts able to comprehend business process models–study insights involving novices and experts. arXiv preprint arXiv:2107.02030.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.

Zhou, C., Zhang, D., Chen, D., & Liu, C. (2023). Business process complexity measurement: A systematic literature review. *IEEE Access*.

Zimoch, M., Mohring, T., Pryss, R., Probst, T., Schlee, W., & Reichert, M. (2017). Using insights from cognitive neuroscience to investigate the effects of event-driven process chains on process model comprehension. In *International conference on business process management* (pp. 446–459). Springer.

Zimoch, M., Pryss, R., Schobel, J., & Reichert, M. (2017). Eye tracking experiments on process model comprehension: lessons learned. In *Enterprise, business-process and information systems modeling*.

Zugal, S. (2013). *Applying cognitive psychology for improving the creation, understanding and maintenance of business process models* (Ph.D. thesis), University of Innsbruck.