Information Technology and Management (2021) 22:99–131
https://doi.org/10.1007/s10799-021-00327-7

Influence of diagram layout and scrolling on understandability of BPMN processes: an eye tracking experiment with BPMN diagrams

Daniel Lübke[1] · **Maike Ahrens**[1] · **Kurt Schneider**[1]

Accepted: 10 March 2021 / Published online: 10 April 2021
© The Author(s) 2021

Abstract
Business process modeling is an important activity for developing software systems—especially within digitization projects and when realizing digital business models. Specifying requirements and building executable workflows is often done by using BPMN 2.0 process models. Although there are several style guides available for BPMN, e.g., by Silver and Richard (BPMN method and style, vol 2, Cody-Cassidy Press, Aptos, 2009), there has not been much empirical research done into the consequences of the diagram layout. In particular, layouts that require scrolling have not been investigated yet. The aim of this research is to establish layout guidelines for business process modeling that help business process modelers to create more understandable business process diagrams. For establishing benefits and penalties of different layouts, a controlled eye tracking experiment was conducted, in which data of 21 professional software developers was used. Our results show that horizontal layouts are less demanding and that as many diagram elements as possible should be put on the initially visible screen area because such diagram elements are viewed more often and longer. Additionally, diagram elements related to the reader's task are read more often than those not relevant to the task. BPMN modelers should favor a horizontal layout and use a more complex snake or multi-line layout whenever the diagrams are too large to fit on one page in order to support BPMN model comprehension.

Keywords Experiment · Eye tracking · Layout · Scrolling · BPMN

1 Motivation

Business process model and notation (BPMN) is an OMG standard [39] that offers a powerful modeling language for defining, documenting, and executing business processes.

Because of its expressiveness, BPMN is often used in process-related software projects (e.g., [29, 55, 62]). With the increasing demand of fully digitized solutions, more and more software systems support or completely control business processes and contain parts that are implemented as workflows. BPMN offers formal, token-based execution semantics and is therefore a single language to specify (analytical modeling) and to execute business processes (executable modeling).

Research has been done into quality attributes of BPMN layout (e.g., modeling guidelines by White [66], layout directions by Figl and Strembeck [19] and spacing of BPMN elements by Scholz and Lübke [50]). However, there are still open questions left, which require empirical studies into the modeling and understandability of BPMN processes.

Understandability is one key quality attribute of BPMN diagrams, especially during the requirements phase in software projects: Technical and non-technical stakeholders have to read, validate, and review complex business processes that will guide further software development. The larger BPMN diagrams get, the more scrolling is required when reading a model on a computer screen, e.g., because the model is made available via an online repository or intranet site. Scrolling may represent an additional barrier when trying to understand a diagram and hence impact comprehension. However, to the best of our knowledge no research into layout options for BPMN diagrams that can avoid scrolling has been done yet.

✉ Daniel Lübke
daniel.luebke@inf.uni-hannover.de

Maike Ahrens
maike.ahrens@inf.uni-hannover.de

Kurt Schneider
kurt.schneider@inf.uni-hannover.de

[1] FG Software Engineering, Leibniz Universität Hannover, Welfengarten 1, 30167 Hannover, Germany

In order to narrow this gap, we conducted a controlled eye tracking experiment with 21 professional software developers. The goal of this experiment is to investigate and characterize the impact of scrolling and diagram layout on understandability. We analyzed different layout directions (horizontal vs. vertical), scrolling (diagram requires scrolling or not), and two two-dimensional layouts (layouts that break horizontally to a new line of BPMN shapes) that can be used to layout a large diagram on a single page.

This article is structured according to suggestions by Wohlin et al. [67, p. 154] as follows: Related Work is presented next in Sect. 2. Section 3 presents the design of the experiment including the research questions, which are derived by using GQM. Section 4 presents the execution of the experiment while in Sect. 5 the plain results and statistical analysis are presented. These are interpreted and put into context in Sect. 6. Finally, we give a conclusion, as well as an outlook on possible future work in Sect. 7.

1.1 Nomenclature of BPMN

BPMN clearly separates between the process model and its visualization, i.e., the diagram(s) representing the model. Because we are concerned with the layout of BPMN, we use the term BPMN diagrams where appropriate. Also, BPMN knows the concept of "activities". However, in contrast to other modeling languages, activity is a super class for atomic and structured activities. BPMN's "task" is the element which represents an atomic unit of work and is the only activity type used in our models. As such, we use the term task instead of activities in the context of this paper.

2 Related work

Quality of business process models and their visual representations is multi-faceted. Lindland et al. [27] have specified a framework that can be used to categorize different quality aspects of models. They differentiate between syntactic, semantic and pragmatic qualities. This paper is concerned with understandability of BPMN process models. Understandability is the main concern of pragmatic model quality, which "affects how to choose from among the many ways to express a single meaning." [27, p. 47] The same business process can be modeled with the same syntax but with different choices concerning visualization. These different choices, e.g., Schrepfer et al. [51] identifies "good" choices for line crossings, edge bends, symmetry, and locality, can greatly influence the understandability of process models. The overall layout, i.e., the direction of modeling, is also thought to affect understandability [18] but—as stated by Dikici et al.—"we still have limited knowledge on the factors contributing to understandable process models." [11]. Thus, this research aims to add some more pieces to the puzzle of understandability by investigating the impact of modeling direction in BPMN models on understandability closer.

Layout of models in general has been studied for a long time. For example, Purchase et al. published their results with regard to general graph understanding. In experiments [47] they showed that reducing the number of arc crossings helps understandability in UML diagrams. They refined the results [46, 48] and experimentally established a priority list of attributes that make diagrams easier to understand. These attributes are arc crossings, orthogonality, information flow, arc bends, text direction, width of layout, and font type. Based on these findings, Purchase defined a set of layout aesthetics covering most of these aspects [45], which apply to all graph-based layouts.

These points were strengthened by Störrle [56, 57], who conducted eye tracking experiments specifically focused on UML Diagrams including Activity Diagrams. Those experiments showed that a clean layout following a set of recommended guidelines—including those of Purchase et al.—is better understandable than a diagram violating those. Bernstein and Soffer [4] added some more characteristics to these attributes, e.g., model size, number of ending points, change in direction, and alignment, by conducting an initial study on business processes. However, their characteristics are not measurable yet.

During the last years more research into notation specific understandability has been conducted; especially of BPMN and UML Activity Diagrams. BPMN research was specific to syntactic elements and constructs both theoretically (e.g., [20, 36, 37]) and empirically (e.g., [22, 40, 42, 43, 49, 50, 63]). A good summary can be found in [17].

One factor affecting understandability is a diagram's layout [19]. This is important because the BPMN standard itself does not require any layout but merely recommends to either use the left-to-right (horizontal) or top-to-bottom (vertical) layout [39].

It seems that laying out diagrams horizontally left-to-right is prevalent. For example, Silver and Richard [54] and Swiss eGovernment standards [5] require left-to-right layout in their guidelines. Layout algorithms have also been implemented that way: Both Effinger and Decker [14, 15], Kitzmann et al. [25], and Scholz and Lübke [50] layout horizontally.

The layout of BPMN diagrams itself has been studied by Figl and Strembeck [19]. They analyzed the understandability of different flow-directions. While the left-to-right layout scored best in absolute numbers, they could not find significant differences for any direction (left-to-right, right-to-left, top-to-bottom and bottom-to-top). Kretschmann arrived at the same conclusions in his Bachelor's Thesis [26], which used eye tracking in contrast to Figl and Strembeck's study.

While reducing or keeping the size [33, 34, 65] and complexity [68] of a process diagram small is the best way to create well-understandable diagrams, added business complexity will increase the process complexity and thus the diagram complexity. This is especially true with executable business processes that require more details to be added to them.

When diagrams get larger, it is possible to restructure the diagram by changing the model itself and use other/additional elements in order to make it smaller and moving parts somewhere else: For example, BPMN supports horizontal and vertical partitioning by the use of link events and collapsed subprocesses. We found no prior studies regarding the understandability of link events, but Turetken et al. investigated structuring BPMN diagrams with subprocesses and groups [59, 60]. Unfortunately for modelers, they found that using these is significantly worse to understand than leaving the process diagram larger and on one level.

This finding means that other options need to be investigated: For example, it is also possible to deal with larger diagrams by changing their layout to make them fit on one page or screen. We also found no prior research investigating these options, which is what we present in this paper.

3 Experimental design

3.1 Theoretical research question

Understandability is a key attribute of BPMN models. Poorly comprehensible BPMN models lead to delays or misunderstandings and errors. However, there are limited options to pick from, and a few assumptions on how to layout large BPMN diagrams. However, are they myths or valid?

There seems to be an assumption that horizontal layouts are preferable over vertical ones [18]. Although this assumption is backed by plausible arguments (scrolling economy), it has not been substantiated with respect to BPMN. Furthermore, for large BPMN diagrams, there are additional variants such as "snake design" that could be considered.

Theoretical Research Question: Are there objective indications supporting the widely held preference of horizontal layout over others with regards to large BPMN diagrams?

Because in the context of BPMN process execution, diagrams tend to get larger, we focus on the layouts of large diagrams. Due to this context the use of computer screens is necessary because no paper printouts can be used to design executable BPMN.

Improving a notation can only succeed in small and detailed steps, given the constraints imposed by syntax and current pragmatics of use. However, each improvement step has substantial impact on the use and benefit of the notation. Making best use of display space ("real estate") is an important aspect. Because BPMN offers a free choice of flow directions, designers must choose how to layout the business process diagram. Is there one that is objectively preferable over the others?

On a theoretical level, we wanted to confirm or reject the widely shared preference for horizontal designs, from an objective aspect of readers' points of view. We apply eye tracking to measure sub-conscious reader activity objectively rather than relying on subjective opinion statements in hindsight that are often biased by assumptions.

3.2 Goals and hypotheses

We used the Goal-Question-Metric Paradigm (GQM) [2] for planning our experiment. GQM was initially invented to guide metrication and evaluation of software process improvement activities [61].

GQM was invented to overcome a frequent problem of dashboards and metrication programs: Metrics were selected based on their availability rather than their relevance. While it is straight-forward to select well-known metrics that are easy to apply, such an approach, however, is not focused on reaching the intended goal behind the measurement. For that reason, Basili suggested to select those metrics that best serve a set of goals rather than those measures. Therefore, defining goals precedes definition of questions and selection of metrics that are suited to answer the questions. This shift in attitude turns evaluation from an unfocused collection process to a goal-oriented and systematic process of planning the measurement, selecting or maybe even constructing new metrics for that purpose.

When used in industry, GQM plans often include subjective aspects, e.g., customer satisfaction measured by a questionnaire. In most industrial applications, measurement programs are tools for gaining insights and plausible hints on the usefulness of an improvement activity. In research, however, an ultimate purpose of measurement is evaluation. Scientific evaluation may use qualitative data, but quantitative data can be better analyzed statistically. The level of

statistical significance is an important indicator of effect validity.

The goal-oriented approach of GQM can be highly instrumental in connecting improvement or measurement intentions (goals) with respective questions and metrics. As Nick and Tautz [38] and Brill et al. [7] demonstrated, this extended use of GQM in guiding research can be very helpful. Ahrens et al. [1] used GQM for focusing an eye tracking experiment in software engineering before. We followed the original GQM process by defining goals, questions, and metrics in that order.

3.2.1 Goal and metrics

Our goal of this experiment was

For the purpose of *understanding* with regard to the quality aspect of *understandability* of the object of *a large BPMN diagram* from the viewpoint of *a reader of that model*.

Research questions represent the GQM questions in this setting. They are organized in alignment with the two subgoals of our main goal:

Diagram layouts RQ1–RQ7 are concerned with the impact of the general diagram layout on understandability. We distinguish horizontal with and without scrolling, vertical with and without scrolling, horizontal/snake, and horizontal/multi-line layouts as defined in Sect. 3.2.2.
Task attributes Depending on the diagram layout, tasks (like any other BPMN element) can be moved off the first screen and require scrolling to be seen and tasks are asked for in questions or not. Depending on these attributes, tasks might be read more intensively.

The research questions are made operational by defining hypotheses and metrics. The whole GQM tree is shown in Fig. 1. In contrast to most research design descriptions using GQM, we will describe the metrics first. Because the same set of metrics is used for most research questions, we later refer to the metrics from the research questions in order to give a more compact and easier documentation of our research design.

Answer time The required duration for answering four questions for each diagram. We compare these times for two diagrams with contrasting layout options.
Error rate The number of errors made while answering the questions for a given diagram. Because we ask four questions for each diagram, this metric can be any number from 0 to 4. Other experiments use the inversion, i.e., the number of correct answers also called task effectiveness (e.g., the experiment done by Turetken et al. [60]).
Task efficiency Task efficiency is the ratio of correct answers per required time [34]. Within our experiment this metric will be measured as correct answers/minute.
Subjective preference We ask participants to give their subjective preference between two layout options or a Likert-scale level for a quality property. These questions are asked cumulatively at the end of the experiment in an online survey.
Fixation count (diagram and task) Fixations are the stabilization of the eye on an object of interest [52]. In this context the fixation count is the number of eye fixations on elements in the BPMN diagram. Goldberg et al. report that a higher number of fixations indicates a less efficient search for finding relevant information on a stimulus [21].

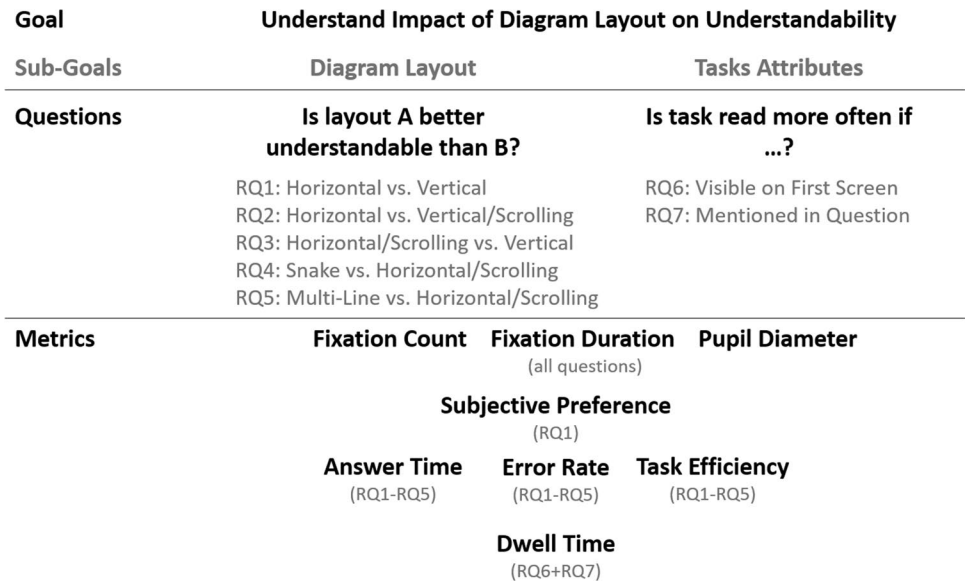

Fig. 1 GQM tree of our research design

It is also used as a measure for visual effort [52]. We analyze fixation counts on two levels: on the one hand for the whole diagram for answering questions related to the layout and on the other hand for specific tasks for answering task-related questions.

Fixation duration (diagram and task) The average duration of fixations while looking at a diagram, given in milliseconds. Djamasbi et al. [12, 13] demonstrated that the length of fixations positively correlates with cognitive load. Therefore, we use this measure as an indirect measure for cognitive load [41]. Like the fixation count metric we analyze fixation duration on both the diagram and task level.

Dwell time The total, aggregated time spent looking on a certain diagram element, i.e., the sum of all fixations during a single visit of an area of interest [8]. We use dwell time as a synonym of total dwell time, i.e., the sum of all dwell times on an element, which describes how much attention is paid to it.

Pupil diameter (diagram and task) The diameter of the pupil measured in millimeter and averaged for both eyes. The pupil widens as a response to low light conditions [52]. Studies showed that it also happens during complex cognitive tasks [23, 44, 53]. Wahn et al. conducted an experiment showing that the pupil size scales with attentional load and task experience, stating that the pupil dilates more the less experienced the subject is in the given task [64]. In our experiment we use the pupil diameter as an additional measure for cognitive load. Like fixation count and fixation duration we analyze pupil diameter both on diagram and task level.

By using eye tracking it is possible to measure indicators of cognitive load. This allows to objectify cognitive load, i.e., not measure it as "perceived understandability" but as "objectively measured understandability" as classified by Dikici et al. [11].

In the following sections the research questions are described in more detail with our hypothesis. This structure will be picked up in Sect. 6, which will present our interpretation based on the aggregation of our measurements and hypothesis test results.

3.2.2 Research questions regarding diagram layout

BPMN modelers have different options for structuring the diagram. They can arrange (a) the control-flow horizontally or vertically (b) in case of large diagrams have the model reader scroll to see all elements or not, and (c) use a bending strategy (no bending, snake or multi-line as illustrated in Fig. 2) to show all diagram elements on one screen in order to avoid scrolling.

In principle, this results in a factorial design of $n := 2 \times 3 \times 3 = 18$ different diagram types. There are 306 $(18^2 - 18)$ total combinations of those layouts. For practical reasons it was necessary to narrow the number of examined diagram types down for reducing the time spent by our subjects in front of the eye tracker. In order to choose which diagram layouts are the most relevant in practice, we analyzed two sources: BPMN models from the industrial project Terravis [3] and the analysis of flow directions on GitHub repositories by Lübke and Wutke [30].

The analysis of the BPMN models of the industrial project Terravis [28] showed that mostly horizontal modeling was used. When scrolling could be avoided and the diagram contained no participants (i.e., pools/(swim-)lanes), multi-line bending was applied, and otherwise the diagrams needed to be scrolled (e.g., when modeling test cases [29] a strict left-to-right order without bending was followed). Sometimes a vertical diagram layout with or without scrolling but without bending was used.

The study of BPMN models on GitHub by Lübke and Wutke [30] showed that 78.21% of the diagrams without pools were layouted horizontally without bending, 9.71% were layouted vertically without bending, 2.19% had a

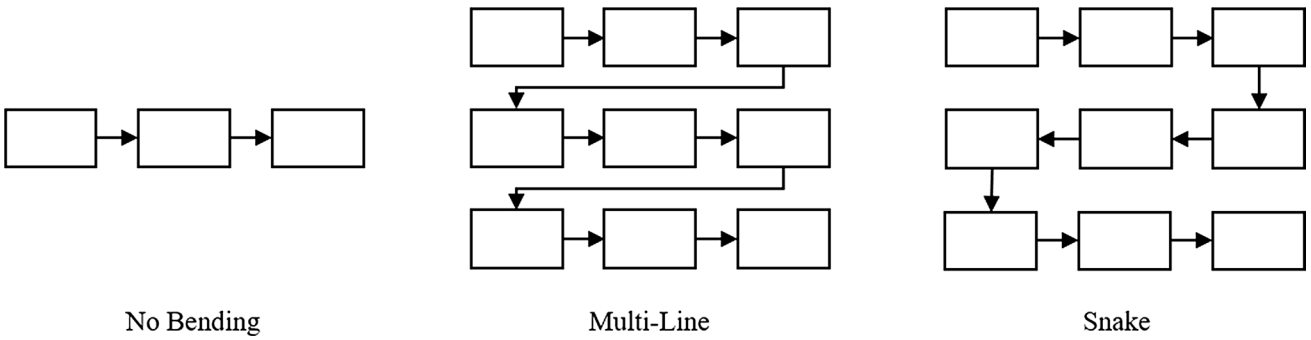

Fig. 2 Different bending strategies used for the horizontally laid out BPMN diagrams

horizontal layout with snake bendings, 0.41% had a horizontal layout with multi-line bendings, 0.24% had a vertical layout with snake bendings, and only 0.12% had a vertical layout with multi-line bending (9.12% were classified as "other layout") The necessity to scroll was not investigated, which is probably not possible, because the environment (e.g., screen size and resolution), which they are viewed in, is unknown.

Based on this data we decided to select only the most relevant combinations, i.e., horizontal and vertical layouts with and without scrolling, and different bending strategies for horizontal layouts. Regarding the understandability of these layouts we formulated research questions RQ1 to RQ5. In this context, we define understandability three-fold: Its first component is the time required to conduct a task, the second component is the required cognitive load, and the third component is how many errors are made.

With these metrics and definition of understandability in mind, the research questions are presented in the following. Each research question compares two diagram layouts. A summary of these comparisons is given in Table 1. Operational hypothesis are given in the "Appendix" for shortening the descriptions of the research questions below.

RQ1: Is a "horizontal" or a "vertical" layout (both without scrolling) of BPMN diagrams better understandable?

Hypothesis: A horizontal layout is better as long as no horizontal scrolling is required. We suppose that the layout direction preferred by the BPMN community (e.g., see existing modeling guidelines [5, 66]), i.e., horizontal layout, is best suited for modeling BPMN diagrams. This hypothesis is also strengthened by the (non-significant) findings of Figl and Strembeck [19]. If any of those two layouts should be superior to the other, horizontal layout scored a bit better so far in experiments. Furthermore, Tufte has shown that people relate the x-axis with time [58]—at least in Western cultures but not necessarily in Asian cultures [6], which strengthens our assumption for preferring the horizontal layout in BPMN diagrams.

Metrics: Answer time, error rate, task efficiency, preference of the subjects indicated in a questionnaire, fixation count, fixation duration, pupil diameter

Operationalized Hypothesis: see Sect. A.1

RQ2: Is a "horizontal" or a "vertical layout with scrolling" better understandable?

Hypothesis: We suppose that a diagram that does not need to be scrolled is easier and especially faster to understand. Scrolling involves user interaction with the system and memorization of off-page content and thus will slow down model reading.

Metrics: Answer time, error rate, task efficiency, fixation count, fixation duration, pupil diameter

Operationalized Hypothesis: see Sect. A.2

RQ3: Is a "vertical" or a "horizontal layout with scrolling" better understandable?

Hypothesis: Although the layout direction is different than with the previous research question, we also suppose in this constellation that every diagram that does not need to be scrolled is easier and especially faster to understand. The reasoning is the same as with RQ2. However, it must be noted that even though the layout directions in RQ2 and RQ3 differ, in both cases the scrolling is vertical, i.e., the mouse wheel could be used.

Metrics: Answer time, error rate, task efficiency, preference of the subjects indicated in a questionnaire, fixation count, fixation duration, pupil diameter

Operationalized Hypothesis: see Sect. A.3

RQ4: Is a horizontal snake layout better understandable than a horizontal layout with scrolling?

Hypothesis: In general we suppose that scrolling is disadvantageous to understandability. Hence, we suppose that a snake layout, which eliminates the need for scrolling, is better to understand, although part of the diagram is arranged in right-to-left direction with snake layouts on every second row.

Metrics: Answer time, error rate, task efficiency, fixation count, fixation duration, pupil diameter

Operationalized Hypothesis: see Sect. A.4

RQ5: Is a horizontal multi-line layout better understandable than a horizontal layout with scrolling?

Hypothesis: We suppose that a multi-line layout, which eliminates the need for scrolling, is better to understand because we generally suppose that scrolling is disadvantageous to understandability.

Metrics: Answer time, error rate, task efficiency, fixation count, fixation duration, pupil diameter

Operationalized Hypothesis: see Sect. A.5

Table 1 Summary of research questions concerning diagram layout

RQ	Layout A	Layout B	Hypothesis for better layout
RQ1	Horizontal	Vertical	Horizontal
RQ2	Horizontal	Vertical/scrolling	Horizontal
RQ3	Vertical	Horizontal/scrolling	Vertical
RQ4	Horizontal/snake	Horizontal/scrolling	Horizontal/snake
RQ5	Horizontal/multi-line	Horizontal/scrolling	Horizontal/multi-line

3.2.3 Research questions regarding reading tasks characteristics

Regardless of the layout of the BPMN diagram, readers might use different strategies for reading and capturing the diagram's contents. Research Questions RQ6 and RQ7 are related to this topic and are presented in this section.

RQ6: Are tasks that are located outside of the initially visible area read less?

Hypothesis: When creating diagrams that require scrolling, some elements are located outside the initially visible area. Since scrolling requires effort on part of the model reader and the reader is only interested in answering his/her questions, we suppose that the initially hidden elements are only looked at if absolutely required to solve a model reader's task. Therefore, the elements located outside the first screen are read less.

Metrics: Dwell time, fixation count, fixation duration, pupil diameter, comparing BPMN tasks located on the initially visible area to those located on the initially not visible area.

Operationalized Hypothesis: see Sect. A.6

RQ7: Are tasks read selectively based on the job of the model reader?

Hypothesis: We suppose that model readers, when presented with a task, will try to locate the elements with the names as they are mentioned in the corresponding question. Other diagram elements will be skipped if possible.

Metrics: Dwell time, fixation count, fixation duration, pupil diameter, comparing BPMN tasks mentioned in the questions to BPMN tasks not mentioned in the questions.

Operationalized Hypothesis: see Sect. A.7

Table 2 Independent, dependent, and extraneous variables of this experiment

Variable	Type	Applies to	Values	Description
Diagram layout	Independent	Diagram	Multi level	The overall layout of the diagram. One of horizontal, vertical
Flow bendness	Independent	Diagram	Multi level	Possible bending strategy for a diagram in order to avoid scrolling. One of none, snake, and multi-line
Scrolling	Independent	Diagram	Boolean	Whether the diagram requires scrolling to view all diagram elements or not
Mentioned in question	Independent	Task	Boolean	Whether the task is referred to in at least one question
Visible on first screen	Independent	Task	Boolean	Whether a task of a scrollable layout is visible on the initial screen to the subject
Fixation count	Dependent	Diagram and task	Count	Number of eye fixations on elements. Goldberg et al. report that a higher number indicates a less efficient search for relevant information [21] and it is also used for measuring the visual effort [52]
Fixation duration	Dependent	Diagram and task	Time/ms	Average duration of fixations; the length of fixations positively correlates with cognitive load [12, 13]
Dwell time	Dependent	Task	Time/ms	Total, aggregated time spent looking on a diagram element [8]
Pupil diameter	Dependent	Diagram and task	Size/mm	Diameter of the pupil averaged for both eyes. Widens during complex cognitive tasks [23]
Subjective preference	Dependent	Diagram	Multi-level	Participant's preference for horizontal left-right or vertical top-down layouts
Answer time	Dependent	Diagram	Time/s	The time required to answer 4 questions for a given diagram
Error rate	Dependent	Diagram	Errors/diagram	Errors being made while answering 4 questions for a given diagram [60]
Task efficiency	Dependent	Diagram	Correct answers/s	Number of correct answers (max. 4) divided by the required time is used to calculate the efficiency for answering the questions. The more correct answers are given and the less time is used, the better the score [34]
Diagram complexity	Extraneous	Diagram	Number	Diagrams vary in size and complexity, e.g., by number of tasks, splits, and joins. We control for this variable by always comparing two diagrams with exactly the same syntactical structure
BPMN experience	Extraneous	Notation	Likert	Participants have different prior knowledge of the BPMN notation making the tasks easier or more difficult to them
UML experience	Extraneous	Notation	Likert	Participants have different prior knowledge of UML activity diagrams which are similar to BPMN. Thus the same reasoning applies

3.3 Design and variables

The experiment is broken down into a series of small experiment segments for every research question that involves diagram layouts. The first part is concerned with RQ1, the second with RQ2, etc. Because every research question regarding overall diagram layouts compares two layouts, a diagram pair D_A/D_B is created for every research question. Further explanations of how these diagrams are created are given in Sect. 3.5.

In order to answer our research questions, independent and extraneous variables as described in Table 2 are used in this experiment: The diagram layout, the bending strategy for the layout, the scrolling of a diagram, the mentioning of a task, and placement of a task on the first screen are independent variables, which we control during diagram creation. Created diagrams and their tasks, gateways, and events are characterized by these attributes. Because bending and scrolling are individually used for avoiding scrolling, they cannot be used in conjunction. In our diagram analysis we also found no instances that used any bending strategy and scrolling at the same time.

However, due to the necessity to make diagrams larger or smaller—dependent on whether they should fit on a single screen, the diagram complexity becomes an extraneous variable. We control this variable by comparing only syntactically identical diagrams which only vary in their layout and thus have the same complexity. Other extraneous variables are prior BPMN and UML Activity Diagram experiences by the subjects. We collect these by using a questionnaire at the end of the experiment.

All metrics as defined in our GQM approach are dependent variables of our experiment (answer time, dwell time, fixation count, fixation time, and pupil diameter).

The experiment is conducted by showing participants a series of diagrams, in which the independent variables are controlled. Each diagram has four questions each. Questions are to be answered with yes or no by checking or unchecking a check-box and are concerned with possible execution orders, e.g., is task A always executed before task B.

We use a counterbalanced, within-group design, which increases the number of data points for statistical analysis while addressing learning effects: On the one hand, because subjects are not split in two unrelated groups, the full number of subjects are available to hypothesis testing. On the other hand, counterbalancing reduces learning effects because subjects are not shown diagrams in the same order so that differences to learning effects should level out. Counterbalancing is done by assigning participants to two groups. The first group is shown diagrams in

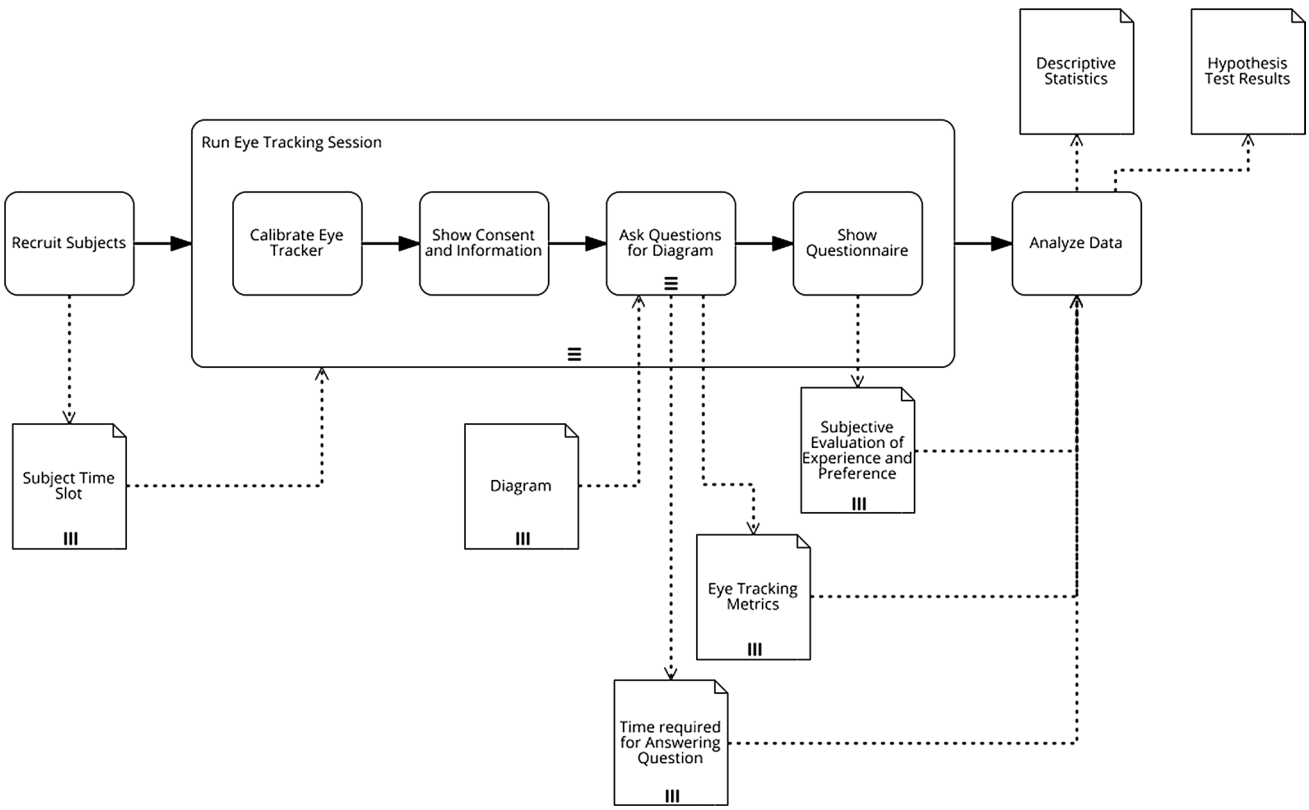

Fig. 3 Experiment process

normal order while the second group sees the diagrams reversed for every diagram pair.

The experiment process as illustrated in Fig. 3 consists of the following steps:

1. *Recruit subjects* By passing a list with reserved time slots around, subjects can choose to participate in the experiment by simply entering their name.
2. *Calibrate the eye tracker* After the participant has entered the room, the eye tracker is calibrated by using the eye tracker supplied software. The process is supervised by the experimenter and feedback is given until the calibration is completed.
3. *Open the Web application and show the participation agreement* The first Web page shows the experiment agreement to the subject and his/her rights to withdraw at any time. Subjects proceed by clicking a "OK" button.
4. *Answer questions for different diagrams* Subjects are then presented with the different diagrams and must answer four questions while their eye movements are being tracked and the times required for answering all questions are recorded.
5. *Complete questionnaire* A questionnaire is presented to the subjects at the end, which asks them for their prior BPMN and UML experience and subjective preferences.
6. *Analyze data* After all subjects have completed the experiment, calculate descriptive statistics and perform hypothesis tests on the collected data.

3.4 Subjects

In the Information Systems domain many experiments are conducted with students (for example, Compeau et al. [9, Table 1] report 36% of studies in the analyzed journals are conducted with students). This frequently raises questions of generalizability [9]. As has been shown by Mendling et al. [32] subject's backgrounds ("Model Viewer Characteristics") are important for their understanding of process models. In order to avoid these issues and because we are interested in larger diagrams typically found in BPMN models in execution projects, we want to recruit our subjects from the pool of professional software developers. We have arranged an agreement to conduct the experiment during a 2-day company training event for recruiting professional software developers.

3.5 Objects

We created a pair of BPMN diagrams for every research question that is concerned with the understandability of

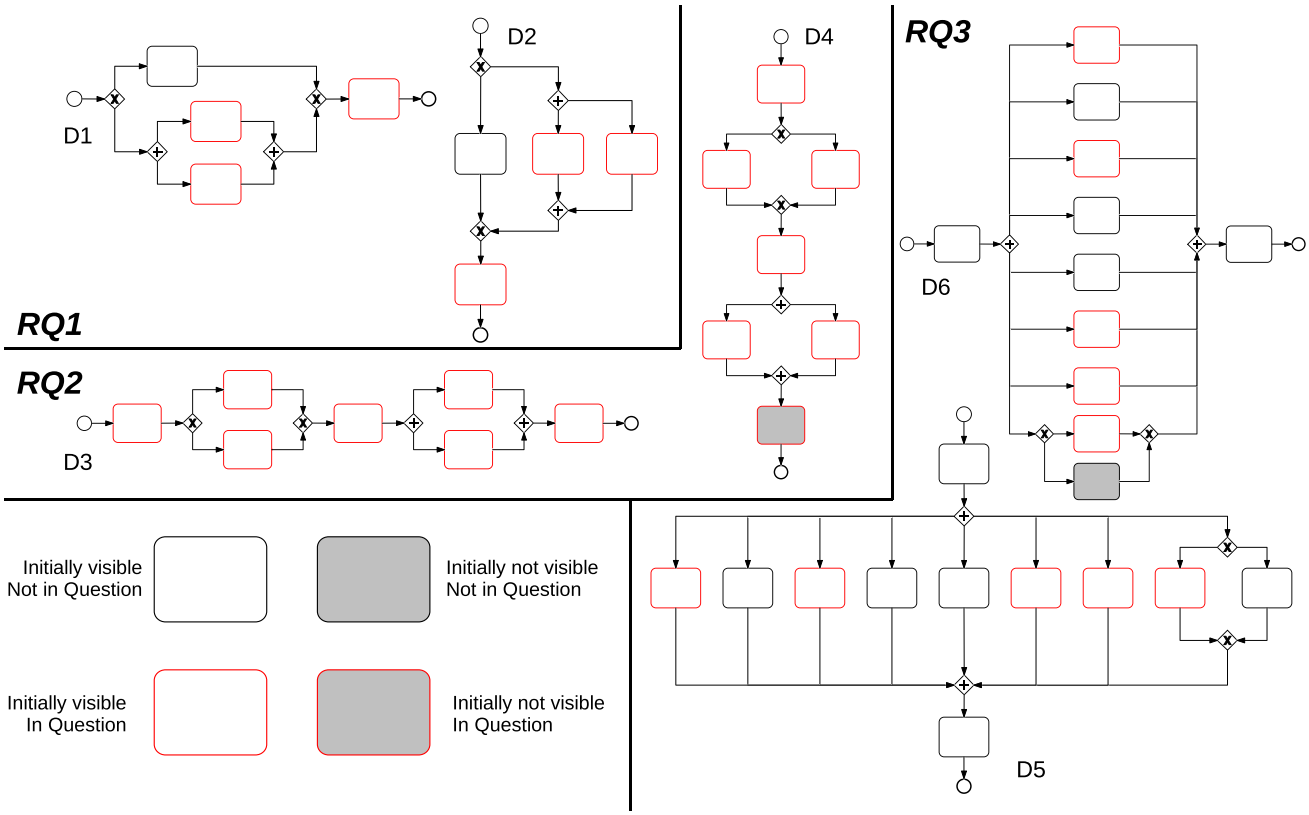

Fig. 4 Diagram layouts used in the experiment (I)

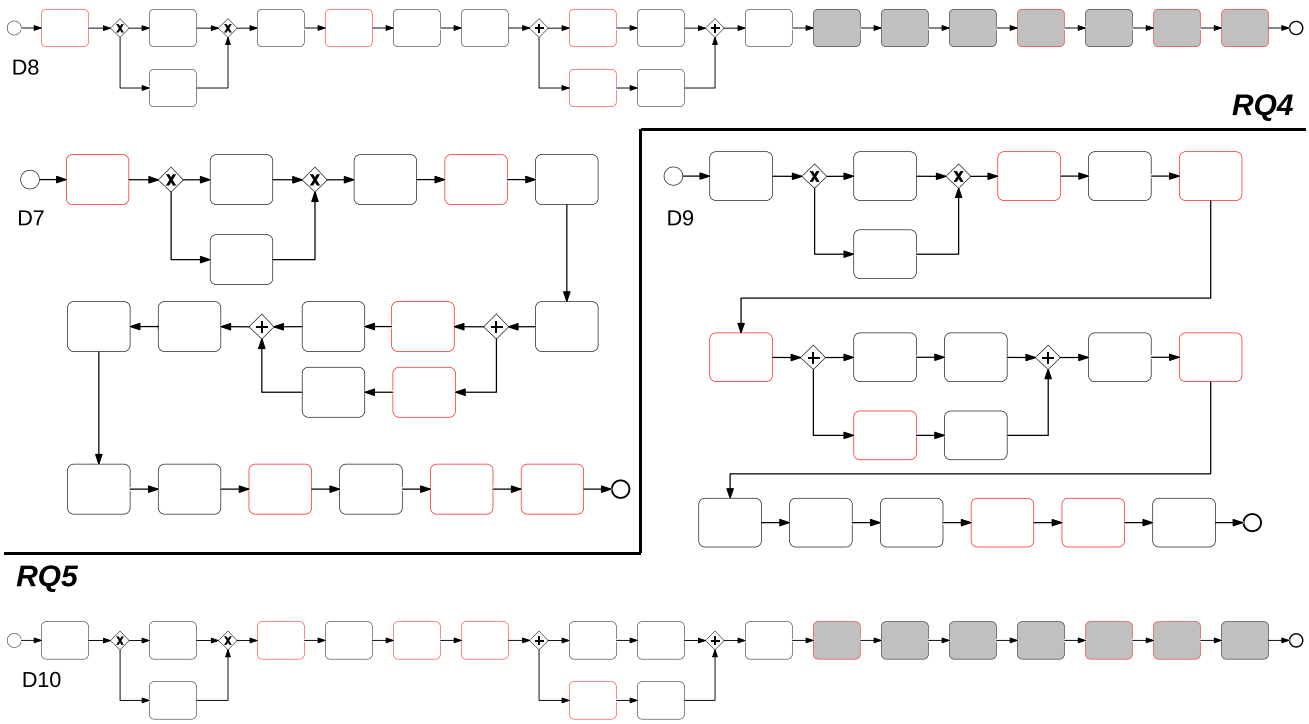

Fig. 5 Diagram layouts used in the experiment (II)

different layouts. Both diagrams show the same BPMN process only with a different layout. Thus, the diagram is created so that the required dimensions in each layout (e.g., for requiring scrolling in one layout but not in the other) are reached.

The diagrams are named D1 to D10. The pairs D1/D2, D3/D4, etc. represent the same model with different layouts. These are shown grouped by research question in Figs. 4 and 5. Tasks in these figures are color-coded: Those that are visible on the first screen have a white background while those that required scrolling to be read have a gray background. Tasks that are not mentioned in a question have a black border while those that are mentioned in a question have a red border.

In order to combat learning effects, subjects are split into two groups and those groups are counter-balanced, i.e., one half of the participants (Group A) is shown the diagrams in ascending order, i.e., D1, D2, ..., D10. The other half (Group B) is presented in an order that reversed the diagrams per research question, i.e., D2, D1, D4, D3, ..., D9. Subjects are assigned randomly to either group dependent on the order of acceptance of participation, i.e., the first registered subject is assigned to group A, the second to group B etc.

Because subjects in the chosen experiment design must answer questions for the same model presented in two different layouts, this would influence the results if the two diagrams were completely the same. This, we created two variants for every diagram in which the labels are different. For instance, there are diagrams D1a and D1b that share the same structure and layout, but the labels denoting the tasks are different. The labels are also changed consistently in the questions, so that participants had to look and search for the same elements in the diagram—just by different labels. The rationale behind the two versions is to mitigate learning effects. These diagrams are used in the two groups, so that the order for the first group was D1a, D2b, D3a, ...and for the second group the order was D2a, D1b, D4a, ...The structure of diagrams D1a and D1b are shown in Fig. 4 in D1 and so on. The concrete diagrams D1a, D1b, D2a, ...are available in the provided materials together with the raw data set at [31].

For all diagrams we use the following modeling guidelines in order to improve consistency and comparability of the diagrams:

– Task names (tasks are represented by rounded rectangles and are semantically similar to UML activities) consist only of a single upper-case character. Figl and Strembeck [19] have shown that the time spent on reading labels correlates with the labels' length. Because we want to measure structural differences, the shortest possible labels are used in order to maximize differences dependent on the layout. Also, abstract labels combat effects of prior domain knowledge: If subjects know a (similar) process, they might answer quicker (because they know) or may make wrong answers (because they know a different process variant). Abstract labels also eradicate intui-

tive understanding of a process due to its context. Thus, subjects must really scan the diagram and understand the syntax.
- Task names are assigned randomly at design time in order to prevent any implication of order (e.g., the question "Is B executed before A" could easily be answered if the tasks were ordered alphabetically).
- All BPMN models conform to BPMN Level 1 as defined by Silver and Richard [54] but do not make full use of the available syntax.
- All diagrams contain exactly one pair of parallel gateways (diamonds with a plus) and one pair of data-driven exclusive or gateways (diamonds with an X). On the one hand, both types of gateways are included in the BPMN Level 1 subset and therefore should be easy enough to understand even by BPMN novices. On the other hand, they add complexity to the diagrams that makes answering the questions not trivial. By using the same numbers of gateways, the control-flow complexity is comparable between all diagrams. All gateways are modeled explicitly for improving understandability [49].
- All diagrams contain as few tasks as possible but as many tasks as necessary for stretching the diagrams to the required dimensions.
- All diagrams contain exactly one none start event and one none end event both represented by circles (single entry, single exit models).
- All diagram elements have the default shape, spacing, colors and font sizes of the Signavio BPMN editor. No zooming is applied. This also implies that all shapes of the same type always have the same size.
- All edges are unlabeled sequence flows, i.e., no conditional sequence flows nor message flows have been used.

3.6 Instrumentation and data collection procedure

The experiment was conducted in a dedicated room not used for anything else at the company's training event. The room had artificial light for improving eye tracking accuracy. Although Turetken et al. [60] have found that BPMN diagrams are better understandable on paper than on screen, our research is concerned with layout options on-screen because in business process execution projects, modelers will work with the models on their computer in order to add all execution details; working with paper models is not feasible in these scenarios. Thus, we use a laptop for all participants that was reserved for the experiment and has a Full HD resolution (1920x1080px) display; the same one is used for all participants.

The experiment was conducted using a small Web application (see Fig. 6) that presents the experiment introduction, agreement of participation, a small BPMN introduction, the questions alongside the diagrams, and the questionnaire with the subjective assessments and preferences. The browser was set to standard zoom and participants were not allowed to change zoom levels. The subjects navigated the Web application by themselves without assistance of an experimenter. The Web

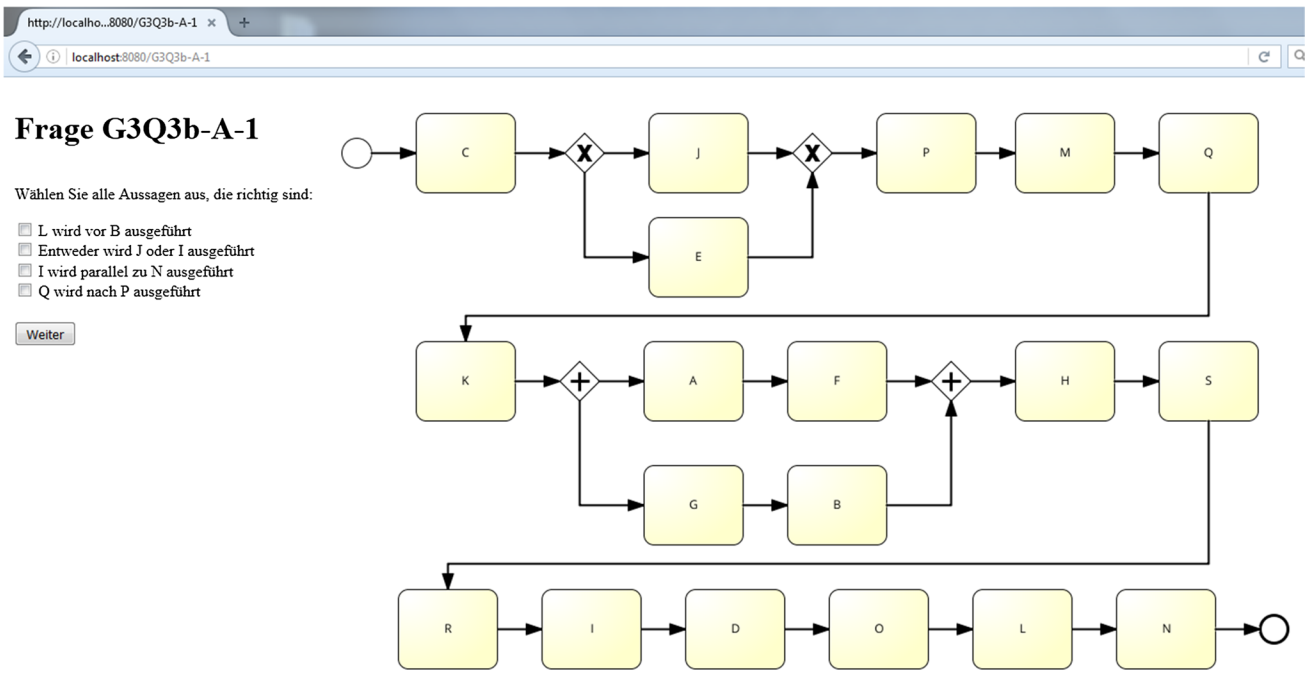

Fig. 6 Screenshot of the web application

application collected the time required for answering all questions for each diagram. This data was written as a log file that was later normalized by using a script and written as a CSV file for further statistical analysis.

For collecting eye tracking data, we use the SMI Red-m remote eye tracker at a sampling rate of 120 Hz. The eye tracker requires calibration, which is guided by the experimenter. A 5-point calibration is used. This is the only task in which the experimenter is allowed to interact with the subjects.

The eye tracking software records the subjects' interactions with the Web application, as well as the eye tracking metrics. In order to be able to compute dwell times of tasks in each diagram, we define areas of interest (AOIs) on each task. To mitigate the effect of small shifts in the recorded positions of gaze points, we define the AOIs of tasks larger than the size of each task itself, but still small enough to prevent overlaps. The software tooling associated with the eye tracker is later used to export a CSV file with eye tracking events.

Our goal was to minimize manual tasks as much as possible and therefore automate the whole data collection process in order to avoid errors made by human experimenters. However, during experiment execution, we noticed that the data export from the eye tracking tool was not correct at all times. In some cases the export files were differing on subsequent exports although no changes or further recordings had been made. We conducted exports and verified them manually by comparing the values in the exported files with the recorded data in the software itself until we finally got a fully correct data export.

3.7 Analysis procedure

After the experiment had been conducted, the data was analyzed. For deciding which hypothesis test to use, a Pearson chi square normality test was performed first on the answer times, fixation count, fixation duration, and pupil diameter variables. If the respective variable is normal, a t-test should be used, otherwise a Wilcoxon hypothesis was used. For comparing data, which is aggregated on a diagram and subject level, paired hypothesis tests were used between the different diagram layouts because a direct comparison of layout A to B is possible per subject. All hypothesis tests were performed two-sided.

3.8 Evaluation of validity

To be able to validate the recorded gaze positions in real time during the experiment, a second screen was used that showed the live view of the Web application window overlapped with the currently tracked gaze point. Thus the experimenter could tell the subject to adjust his or her positioning in front of the screen if required, and therefore prevent loss of tracking. Before finally exporting the recorded eye tracking data, gaze plots of all participants on all diagrams were verified with regard to possible offsets and general precision and quality of tracked gaze points.

In order to validate the Web application including its implementation, logging, and presented diagrams and questions, the experiment was done offline with volunteers that are not employees of the same company as the subjects. Findings were fed back to the Web application prior to the experiment run. Additionally, we performed a dry run with three professional software developers of other companies. Some diagrams have not been positioned correctly, as well as few tasks were mislabeled, which was fixed before executing the experiment.

4 Execution

4.1 Sample

Subjects were recruited at an internal 2-day company training event. The company is specialized in software development and consulting services. Thus, all subjects are software engineering consultants. Participants could voluntarily participate by subscribing themselves into a time schedule. 30 min slots were made available on both days. By subscribing to a time slot, subjects unknowingly assigned themselves to one of the two experiment groups. Participants of even-numbered slots were assigned to group A and participants of odd-numbered slots to group B. No incentives were offered for participation.

In total, 24 software professionals signed up for the experiment. The fastest participant finished the experiment (excluding setup and calibration time) in 6:57 min, the slowest participant took 18:41 min. The average session duration was 10:15 min.

The data of one participant could not be recorded because the person was blind on one eye and the eye tracker could not correctly track the movement of the functioning eye, which was detected during the eye tracker calibration. Therefore, we had to cancel the eye tracking session with this participant. All other subjects completed the whole experiment. When later analyzing the data, two other participants had to be excluded because their data was unusable as described in Sect. 5. This left 21 participants with analyzable eye tracking data.

The characteristics of the sample population is as follows:

Table 3 UML and BPMN experience of subjects

BPMN exp.	UML experience						Total
	1	2	3	4	5	n/a	
1	1	0	0	0	0	0	1
2	1	3	4	1	0	1	10
3	1	4	3	0	0	0	8
4	0	0	1	0	1	0	2
5	0	0	0	0	0	0	0
n/a	0	0	0	0	0	0	0
Total	3	7	8	1	1	1	21

Professional experience/level 1 C-level executive, 16 Senior Consultants, 1 Consultants, and 2 computer science student apprentices, and 1 other staff member (according to the company's classification)
UML and BPMN experience Developers could rate their experience from 1 (no experience) to 5 (very experienced). The most experienced developer rated him- or herself with a 5 for UML experience and a 4 for BPMN experience. The most unexperienced developer rated him- or herself with a 1 for both UML and BPMN experience. The answers to the UML Activity Diagram and BPMN Experience are summarized in Table 3.
Sex The sample consists of 19 male and 3 female subjects.

The random assignment of subjects to their respective group yielded a random distribution of their prior experience. Wilcoxon tests performed for comparing the experience level between both groups resulted in $p = 0.4712$ for UML experience and $p = 0.1546$ for BPMN experience respectively.

4.2 Preparation

The experiment was prepared by announcing it at the opening of training event and passing around the schedule with available time slots. Company employees participating in the event could volunteer to participate by writing their name into a free slot. This list was passed around and there were less available time slots than people interested in participating. However, the time limitation did not allow to offer additional slots. Although no incentives were offered, subjects told us that they were motivated to use an eye tracker as a new technical "toy". The assignment to the groups worked flawlessly and without subjects noticing.

4.3 Validity procedures

Subjects were guided by the Web application through the experiment workflow and answered the questions for the diagrams in the prescribed order. No deviations during the conduction were noticed.

5 Analysis

After the description of how the experiment was conducted, this section will present the results of the experiment, the data, and statistics and hypothesis tests. These will be interpreted later in Sect. 6. The raw data set is available at [31].

5.1 Data set reduction

When analyzing the data of the 23 subjects who took part in the experiment, the eye tracking data of two subjects had to be excluded because the eye tracker recorded only a few tracking points with random offsets to the diagram elements that were not correctable. One of these participants wore glasses. This left 21 participants (19 male, 2 female) with analyzable eye tracking data. We removed the answers for subjective preference and experience from these subjects. This means that we utilize 21 full data-sets only. No further data points were purged from the data set.

5.2 Eye tracking data corrections

We manually analyzed all subjects' gaze plots with regard to validity of the recorded eye tracking data and potentially necessary offset corrections. In six cases (five participants, one or two diagrams each) there was a clearly visible offset in the data in spite of its general quality. Offsets were determined by comparing the pattern of the gaze points with the

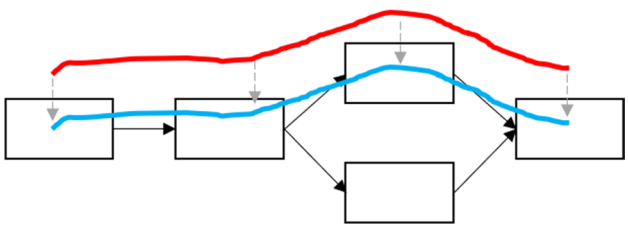

Fig. 7 Eye tracking data offset correction (black: diagram, red: measured eye tracking data, blue: eye tracking data corrected for vertical offset)

structure of the diagram. For those that were unambiguously shifted, we corrected the offset, which was mostly vertical, before further analyzing the data. The principle of this correction is illustrated in Fig. 7.

5.3 Descriptive statistics and hypothesis testing

Within this section, we present the plain results from the measurements in our experiment and the results of the hypothesis tests on this data. We will map those data to the research questions and give an interpretation in Sect. 6.

Before performing any statistics, a test for normality of all variables was performed. All variables except for task efficiency are not normally distributed, because the Pearson chi-square test for normality returned $p < 0.05$ for these variables (Answer Time: $p = 3.695 \times 10^{-7}$, Fixation Count: $p = 3.0431 \times 10^{-7}$, Fixation Duration: $p < 2.22 \times 10^{-16}$, Pupil Diameter: $p < 2.22 \times 10^{-16}$, Error Rate: $p < 2.22 \times 10^{-16}$, Task Efficiency: $p = 0.30245$). Thus, t-tests are performed for task efficiency, and Wilcoxon hypothesis tests are used for all other variables in our analysis.

5.3.1 Answer time

The first metric that we measured is the time required for answering all four questions for each diagram. A box plot of the measured times that the subjects took to complete the four questions, is shown in Fig. 8. Descriptive statistics are presented in Table 4: The quickest answer for all questions was given within 20.0 s (diagram D6) and the longest time taken by one participant was for diagram D9 which took him or her 115.6 s. The quickest diagram to answer on average was diagram D3 which took a mean time of 35.9 s. On average the longest time was required for answering the questions for diagram D8 which took a mean time of 57.9 s.

The pairs of diagrams that showed the same contents but with a different layout, are grouped in Table 4. Every pair tests a hypothesis of a research question. As such we computed the p value by using the Wilcoxon test for paired but not normally distributed data. The table also shows the effect size computed by using Cohen's d and the absolute and relative differences in the mean value (Δ). Here and throughout this article we mark significant p values with stars ((*) < 0.05, (**) < 0.01, (***) < 0.001)) and different effect sizes with crosses ((+) > 0.2, (++) > 0.5, (+++) > 0.8)).

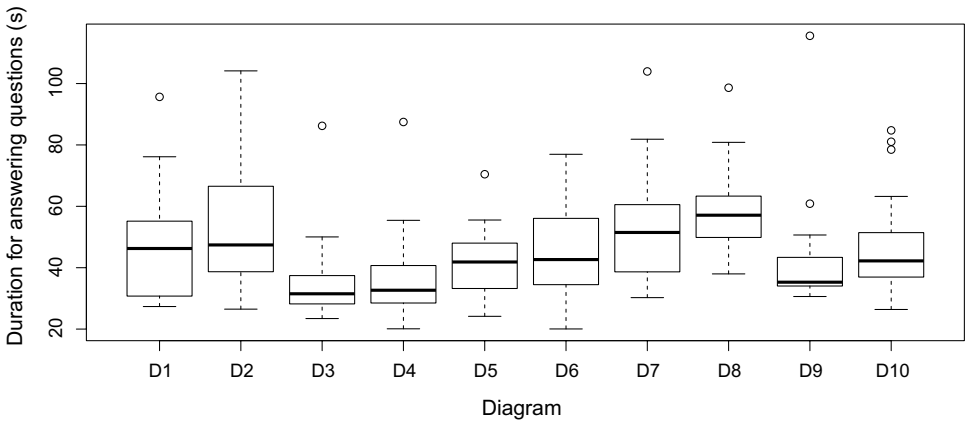

Fig. 8 Boxplot of duration for answering questions (s)

Table 4 Descriptive statistics and hypothesis test results for duration for answering questions (s)

H_1	D.	Min	Mean	Max	p	d	Δ
Horizontal layout is easier to understand than vertical layout	D1	27.3	47.8	95.7	0.228	0.30	5.5
	D2	26.5	53.3	104.1		(+)	11.54%
Horizontal layout is easier to understand than vertical layout with scrolling	D3	23.4	35.1	86.2	0.257	0.20	2.7
	D4	20.1	37.8	87.5		(+)	7.68%
Vertical layout is easier to understand than horizontal layout with scrolling	D5	24.2	41.6	70.5	0.167	0.26	2.9
	D6	20.0	44.5	77.0		(+)	7.00%
Snake layout is easier to understand than horizontal layout with scrolling	D7	30.3	52.7	103.9	0.054	0.31	5.6
	D8	38.0	58.3	98.6		(+)	10.65%
Multi-line layout is easier to understand than horizontal layout with scrolling	D9	30.6	41.9	115.6	0.014	0.31	5.6
	D10	26.4	47.5	84.7	(*)	(+)	13.48%

The most interesting data points are the two diagram pairs with the largest diagrams, which either needed to be scrolled or were arranged with a snake (pair D7/D8) or multi-line layout (pair D9/D10) to fit on one page. In both cases the variant arranged on a single page was answered quicker and yielded significant p values ($p = 0.0484$ for snake layout and $p = 4.85 \times 10^{-3}$ for multi-line layout). Also, both showed a small effect with $d = 0.30$ and $d = 0.39$ respectively.

Additionally, a small effect was shown by the comparison of horizontal and vertical layout (D1/D2: $d = 0.28$), in which answers for the horizontal layout were given quicker. However, the difference was not significant ($p = 0.29 > 0.05$).

In the remaining comparisons of a layout fitting on one page in one direction (e.g., horizontally) and requiring scrolling when arranged differently (e.g., vertically), questions to the non-scrolling diagrams were answered quicker but neither in a significant way nor showing a considerable effect size.

5.3.2 Error rate

Next, we analyzed the errors made while answering the 4 questions per diagram. As such, the maximum number of possible errors per diagram is 4. However, only few mistakes have been made by our subjects. Almost all questions per diagram have been answered without mistake (see Fig. 9): The best-answered diagram was D10, for which no participant made any error. For diagrams D1, D2, D8, and D9 only one participant made one error, while for diagrams D3, D4, D5, and D6 two participants made one error each. The worst answered diagram was D7, for which one participant made one error and two participants made two errors.

All in all, most questions are answered correctly, which can be seen as the single bars for 0 errors in the boxplots. The error rate is low and as such any deviation might be due to chance. There are only differences in the error rate for two diagram pairs, namely D7/D8 (5 errors vs. 1 error in total) and D9/D10 (1 error vs. no error in total) but these are minimal. The largest one is for the diagram pair D7/D8 which has a difference of 4 errors. These small differences are reflected in high p values (see Table 5); accordingly no difference is significant. Because in general the differences are so low and as such the standard deviation is low, the differing diagram pairs show a small effect.

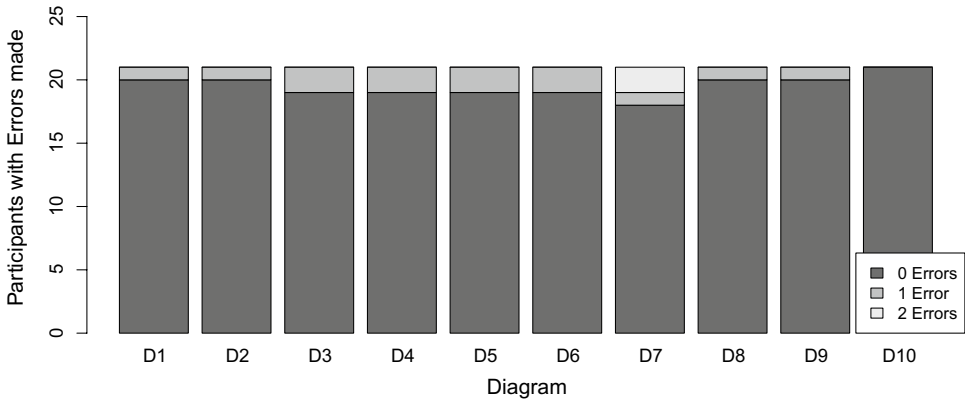

Fig. 9 Boxplot of participants with errors made

Table 5 Descriptive statistics and hypothesis test results for participants with errors made

H_1	D.	Errors made			p	d
		0	1	2		
Horizontal layout is easier to understand than vertical layout	D1	20	1	0	1	0.00
	D2	20	1	0		
Horizontal layout is easier to understand than vertical layout with scrolling	D3	19	2	0	1	0.00
	D4	19	2	0		
Vertical layout is easier to understand than horizontal layout with scrolling	D5	19	2	0	1	0.00
	D6	19	2	0		
Snake layout is easier to understand than horizontal layout with scrolling	D7	18	1	2	0.345	0.30
	D8	20	1	0		(+)
Multi-line layout is easier to understand than horizontal layout with scrolling	D9	20	1	0	1	0.22
	D10	21	0	0		(+)

5.3.3 Task efficiency

The task efficiency score combines the correct answers and the answer times into one metric measured in correct answers per minute. The corresponding boxplots are shown in Fig. 10.

The most ineffective participant answered 1.78 correct answers per minute for diagram D7. The most effective one was a participant answering 11.97 correct answers per minute for diagram D6. The difference in effectiveness between diagrams D5 and D6 is significant ($p = 0.014$) and has a small effect size ($d = 0.44$). Three other diagram pairs also show a small effect size but no significant p values: D1/D2, D7/D8, and D9/D10. The pair D3/D4 has no significant p value nor a demonstrable effect. An overview of the descriptive statistics, p values, and effect sizes is shown in Table 6.

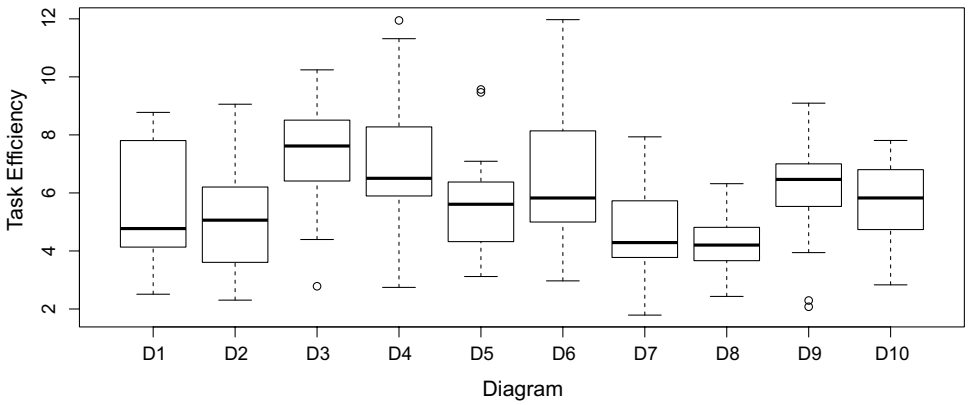

Fig. 10 Boxplot of task efficiency

Table 6 Descriptive statistics and hypothesis test results for task efficiency

H_1	D.	Min	Mean	Max	p	d	Δ
Horizontal layout is easier to understand than vertical layout	D1	2.50883	5.64456	8.77546	0.137	0.25	−0.50279
	D2	2.30483	5.14177	9.05763		(+)	−8.91%
Horizontal layout is easier to understand than vertical layout with scrolling	D3	2.78345	7.24591	10.24197	0.640	0.10	−0.19355
	D4	2.74352	7.05236	11.94565			−2.67%
Vertical layout is easier to understand than horizontal layout with scrolling	D5	3.11891	5.60173	9.56556	0.014	0.44	0.75024
	D6	2.96888	6.35198	11.97246	(*)	(+)	13.39%
Snake layout is easier to understand than horizontal layout with scrolling	D7	1.78888	4.70603	7.93284	0.153	0.26	−0.41843
	D8	2.43314	4.28759	6.31845		(+)	−8.89%
Multi-line layout is easier to understand than horizontal layout with scrolling	D9	2.07693	6.09019	9.09366	0.191	0.25	−0.41866
	D10	2.83189	5.67154	7.80793		(+)	−6.87%

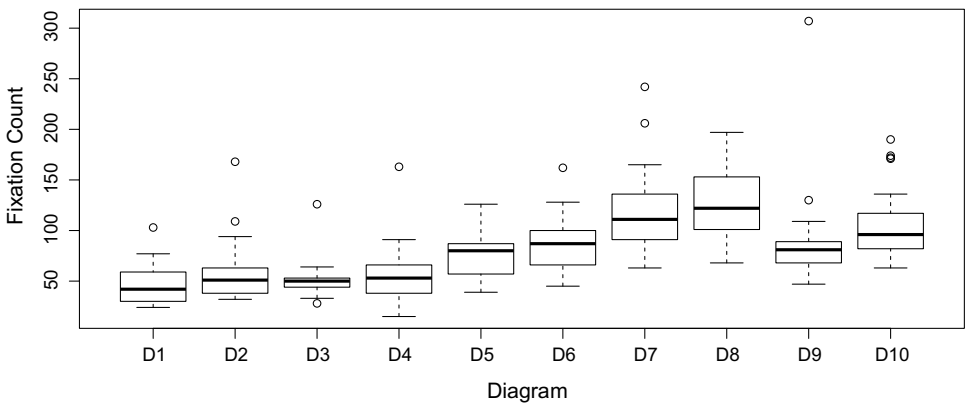

Fig. 11 Boxplot of fixation count

Table 7 Descriptive statistics and hypothesis test results for fixation count

H_1	D.	Min	Mean	Max	*p*	d	Δ
Horizontal layout is easier to understand than vertical layout	D1	24.0	48.0	103.0	0.031	0.52	10.6
	D2	32.0	58.7	168.0	(*)	(++)	22.10%
Horizontal layout is easier to understand than vertical layout with scrolling	D3	28.0	51.2	126.0	0.258	0.28	5.3
	D4	15.0	56.5	163.0		(+)	10.42%
Vertical layout is easier to understand than horizontal layout with scrolling	D5	39.0	76.2	126.0	0.180	0.53	11.6
	D6	45.0	87.8	162.0		(++)	15.18%
Snake layout is easier to understand than horizontal layout with scrolling	D7	63.0	120.7	242.0	0.614	0.15	6.6
	D8	68.0	127.3	197.0			5.44%
Multi-line layout is easier to understand than horizontal layout with scrolling	D9	47.0	91.0	307.0	0.047	0.31	16.4
	D10	63.0	107.3	190.0	(*)	(+)	18.01%

5.3.4 Fixation count

We also measured the fixation counts for the different diagrams and thus different layouts. In total we recorded 17320 fixation events along with the pupil diameter on BPMN diagram elements (tasks, gateways, and events).

Box plots showing the fixation counts for the different diagrams are shown in Fig. 11. The descriptive statistics as well as the results of the hypothesis tests for each diagram pair are shown in Table 7 following the same structure as for the table showing the answer times above.

The diagram that had the fewest fixations on average was D3 (mean 51.2). The one with the most fixations on average was D8 (mean 127.3). The lowest number of fixations on a diagram by a participant was on D4 (15 fixations) and the highest number was by a participant on D9 (307 fixations). On average, the horizontal layout had fewer fixations than the vertical layout (D1/D2: 48.0 vs. 58.7), the non-scrolling layouts had fewer fixations than the scrolling ones (D3/D4: 51.2 vs. 56.5; D5/D6: 76.2 vs. 87.8), and the layouts on one page had fewer fixations than the same diagrams arranged with horizontal scrolling (D7/D8: 120.7 vs. 127.3; D9/D10: 91.0 vs. 107.3).

When conducting Wilcoxon's paired hypothesis tests for the different diagram pairs, a significant difference of mean values is found for the diagram pairs D1/D2 ($p = 0.031$) and D9/D10 ($p = 0.047$). Large effect sizes were found for diagram pairs D1/D2 ($d = 0.52$) and D5/D6 ($d = 0.53$). Small effect sizes were found for D3/D4 ($d = 0.29$) and D9/D10 ($d = 0.26$). Only the diagram pair D7/D8 showed no effect.

5.3.5 Fixation duration

Besides the fixation count we also measured the fixation duration for participants working with different layouts. The box plots for this metric are shown in Fig. 12. The descriptive statistics, as well as the results of the hypothesis tests are shown in Table 8.

The diagram that had the shortest fixation duration on average was D10 (mean 200.2 ms). The one with the longest fixation duration on average was D2 (mean 236.6 ms). The shortest fixation duration on a diagram was by a participant on D4 with 152.2 ms and the longest fixation duration was by a participant on D5 with 312.2 ms.

The horizontal layout of RQ1 had a shorter average fixation duration than the vertical layout (D1/D2: 222.8 vs.

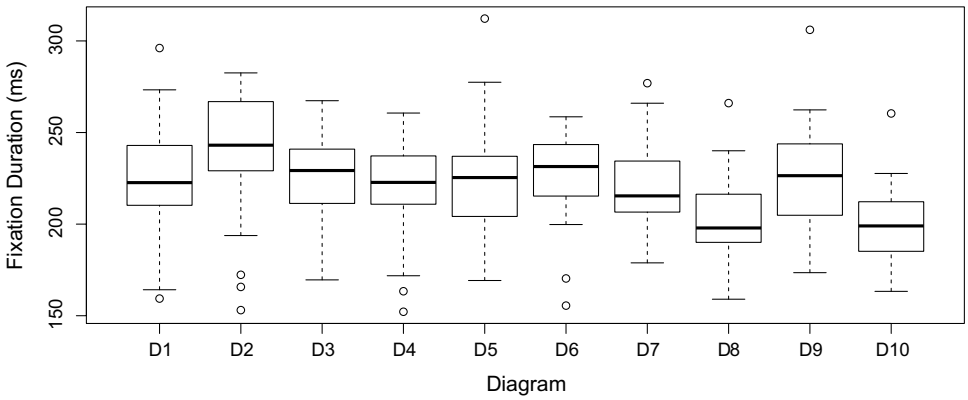

Fig. 12 Boxplot of fixation duration (ms)

Table 8 Descriptive statistics and hypothesis test results for fixation duration (ms)

H_1	D.	Min	Mean	Max	p	d	Δ
Horizontal layout is easier to understand than vertical layout	D1	159.4	222.8	296.1	0.095	0.42	13.8
	D2	153.1	236.6	282.5		(+)	6.21%
Horizontal layout is easier to understand than vertical layout with scrolling	D3	169.6	227.8	267.4	0.373	0.32	−8.6
	D4	152.2	219.2	260.6		(+)	−3.80%
Vertical layout is easier to understand than horizontal layout with scrolling	D5	169.2	224.6	312.2	0.707	0.03	1.0
	D6	155.5	225.6	258.6			0.44%
Snake layout is easier to understand than horizontal layout with scrolling	D7	178.9	219.9	276.9	1.37×10^{-3}	0.70	−16.6
	D8	159.0	203.3	266.1	(**)	(++)	−7.56%
Multi-line layout is easier to understand than horizontal layout with scrolling	D9	173.5	223.7	306.1	2.92×10^{-4}	0.73	−23.5
	D10	163.3	200.2	260.4	(***)	(++)	−10.50%

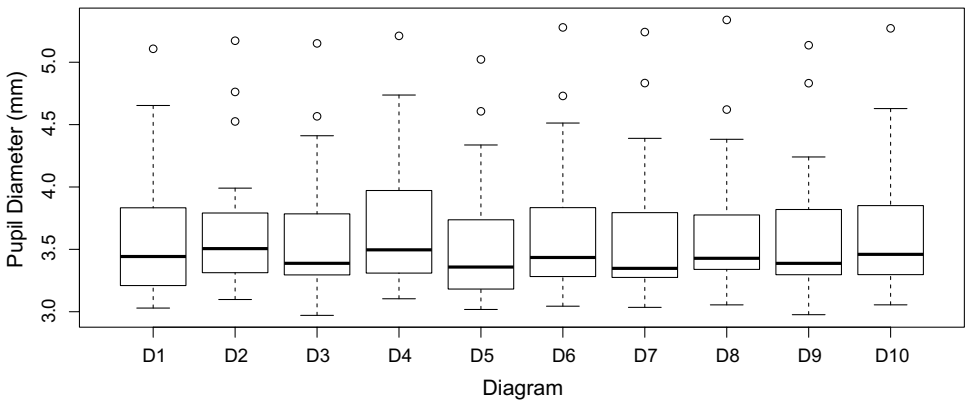

Fig. 13 Boxplot of pupil diameter (mm)

Table 9 Descriptive statistics and hypothesis test results for pupil diameter (mm)

H_1	D.	Min	Mean	Max	p	d	Δ
Horizontal layout is easier to understand than vertical layout	D1	3.03	3.63	5.11	0.038	0.07	0.04
	D2	3.10	3.67	5.17	(*)		1.03%
Horizontal layout is easier to understand than vertical layout with scrolling	D3	2.97	3.58	5.15	6.07×10^{-4}	0.17	0.09
	D4	3.10	3.67	5.21	(***)		2.55%
Vertical layout is easier to understand than horizontal layout with scrolling	D5	3.02	3.55	5.02	1.33×10^{-5}	0.18	0.09
	D6	3.04	3.65	5.28	(***)		2.65%
Snake layout is easier to understand than horizontal layout with scrolling	D7	3.03	3.60	5.24	0.017	0.06	0.03
	D8	3.05	3.63	5.34	(*)		0.88%
Multi-line layout is easier to understand than horizontal layout with scrolling	D9	2.98	3.61	5.14	0.494	0.01	0.01
	D10	3.06	3.62	5.27			0.21%

236.6) and the non-scrolling variants had a longer average fixation duration than the scrolling alternatives (D3/D4: 227.8 vs. 219.2, D5/D6: 224.6 vs. 225.6, D7/D8: 219.9 vs. 203.3, D9/D10: 223.7 vs. 200.2). The difference between large horizontal scrolling layouts vs. the horizontal snake and multi-line layouts showed significant differences and medium effect sizes (D7/D8: $p = 1.37 \times 10^{-3}, d = 0.70$; D9/D10: $p = 2.92 \times 10^{-4}, d = 0.73$). The non-scrolling horizontal layout compared with a non-scrolling vertical layout as well as with a scrolling vertical layout showed a small effect size (D1/D2: $d = 0.42$; D3/D4: $d = 0.32$).

5.3.6 Pupil diameter

Furthermore, we analyzed the pupil diameter measurements. The box plots for this metric are shown in Fig. 13 and the detailed statistics including hypothesis test results are shown in Table 9.

The diagram that showed the smallest pupil diameter on average was D5 with 3.55 mm. The one with the largest pupil diameter was D4 with 3.67 mm. The smallest pupil diameter by a participant was measured on D9 with 2.98 mm. The largest pupil diameter was encountered on D8 measuring 5.34 mm.

Pupil diameters on the horizontal layout of D1 were significantly smaller than on the corresponding vertical layout D2 ($p = 0.038$), however with a negligible effect size ($d = 0.07$, $\Delta = 1.03\%$). The non-scrolling layouts D3 and D5 were viewed with highly significantly smaller pupil diameters than their counterparts D4 and D6 (D3/D4: $= 6.07 \times 10^{-4}$, D5/D6: $p = 1.33 \times 10^{-5}$), also with negligible effect sizes (D3/D4: $d = 0.17$, $\Delta = 2.55\%$; D5/D6: $d = 0.18$, $\Delta = 2.65\%$). The snake layout of D8 led to a significantly smaller pupil diameter than in D7 ($p = 0.017$), but with a negligible effect size. Finally, D9 and D10 had no significant difference in pupil diameter and the effect was nearly zero.

5.3.7 Dwell time of tasks

As shown in the previous sections, we measured the metrics defined via GQM for comparing diagram layouts. For answering the last two research questions, we need to break these metrics down to the BPMN task level.

For directly answering the research questions, we did this separately on the basis of the two factors "visible on first page" and "mentioned in question". For analyzing the combinations, we additionally created four subsets of tasks based on the combination of these two attributes.

The first metric to be broken down is the dwell time, i.e., the time spent by each subject on each BPMN task. The box plots for the distributions of the dwell time are shown in Fig. 14 for single attributes and two attributes combined.

For analyzing the different subsets of tasks clustered by both attributes, we used the non-parametric Kruskal–Wallis hypothesis test in order to determine whether the subsets are differing in their means. The test indicated that the means within the subsets are not equal for the dwell time ($p = 6.036\,58 \times 10^{-16}$). Consequently, we conducted pairwise unpaired Wilcoxon hypothesis tests for comparing the tasks with different attributes with each other. The results are summarized in Table 10.

At first, we compared tasks that are initially visible with those that are not visible without scrolling (see Table 10). Tasks that were not visible on the initial screen were looked at highly significantly shorter ($p = 1.87 \times 10^{-8}$) with a small effect size ($d = 0.32$), which resembles a difference of $\Delta = -22.33\%$ on average. Also, tasks that are not mentioned in one of the questions are also looked at highly significantly shorter $p < 2.0 \times 10^{-16}$ and with a small effect of $d = 0.36$, which accounts to a difference of $\Delta = -26.51\%$ on average.

The largest difference is between tasks that are both visible on the first page and mentioned in a question compared to tasks that are not initially visible and that are not mentioned in a question. The difference of dwell time on average is $\Delta = -40.72\%$ with a very high significance ($p < 2.0 \times 10^{-16}$) and a very large effect size ($d = 0.87$). In general, tasks that combine more "positive" attributes, i.e., being visible on the first page or being mentioned in a question, are looked at longer. This is expressed in the statistics

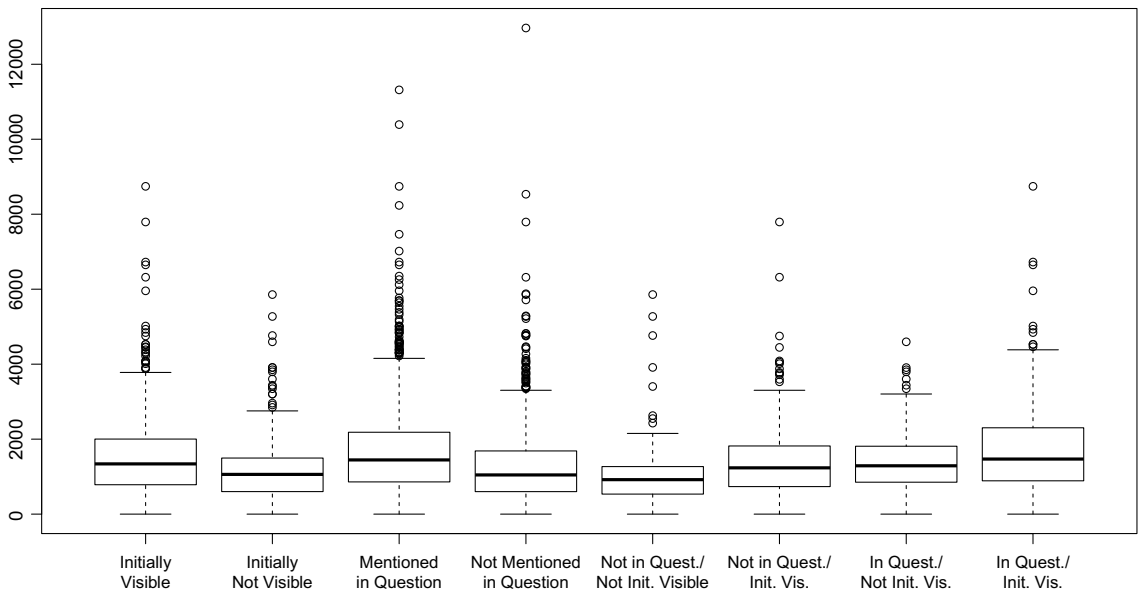

Fig. 14 Dwell time (ms) for BPMN activities by all attributes

Table 10 Descriptive statistics and hypothesis test results for subsets of tasks (dwell time in ms)

Task attr.	Min.	Med.	Mean	Max.	p	d	Δ
Initially visible	0.00	1339.74	1511.80	8743.75	1.87×10^{-8}	0.32	−337.57
Initially not visible	0.00	1059.62	1174.23	5857.06	(***)	(+)	−22.33%
In question	0.00	1445.62	1710.61	11316.62	$< 2. \times 10^{-16}$	0.36	−453.48
Not in question	0.00	1045.58	1257.14	12968.51	(***)	(+)	−26.51%
Not in quest./init. vis.	0.00	1234.79	1376.19	7793.40	1.72×10^{-8}	0.47	−371.15
Not in quest./not init. vis.	0.00	919.36	1005.04	5857.06	(***)	(+)	−26.97%
In quest./not init. vis.	0.00	1289.24	1456.23	4597.02	6.11×10^{-8}	0.57	−451.19
Not in quest./not init. vis.	0.00	919.36	1005.04	5857.06	(***)	(++)	−30.98%
In quest./init. vis.	0.00	1469.05	1695.27	8743.75	$< 2. \times 10^{-16}$	0.87	−690.23
Not in quest./not init. vis.	0.00	919.36	1005.04	5857.06	(***)	(+++)	−40.72%
In quest./not init. vis.	0.00	1289.24	1456.23	4597.02	0.235	0.09	−80.03
Not in quest./init. vis.	0.00	1234.79	1376.19	7793.40			−5.50%
In quest./init. vis.	0.00	1469.05	1695.27	8743.75	5.38×10^{-5}	0.34	−319.08
Not in quest./init. vis.	0.00	1234.79	1376.19	7793.40	(***)	(+)	−18.82%
In quest./init. vis.	0.00	1469.05	1695.27	8743.75	0.075	0.27	−239.04
In quest./not init. vis.	0.00	1289.24	1456.23	4597.02		(+)	−14.10%

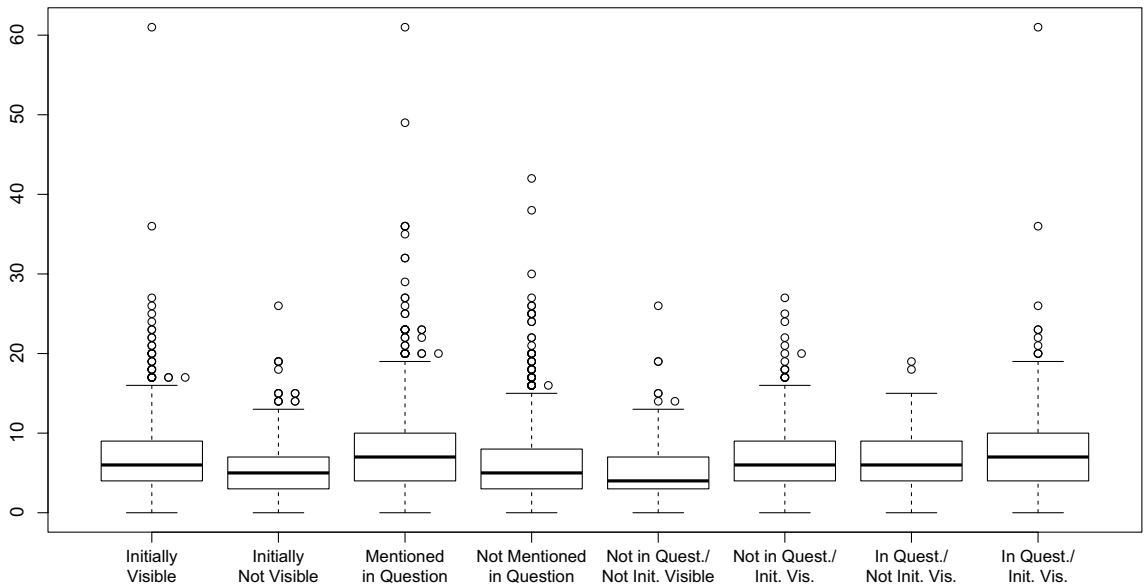

Fig. 15 Fixation count for BPMN tasks by all attributes

by very small p values and small to medium effect sizes. The only exception is when comparing tasks that are both mentioned in a question but one set is being visible on the first page while the other set is not visible on the first page. In this case, no significant difference and no effect is found. Also the combination of tasks initially visible and not mentioned in a question compared to tasks which are not initially visible but mentioned in a question yielded no significant results although a small effect ($d = 0.27$) was found for the absolute difference of $\Delta = -14.10\%$.

5.3.8 Fixation count of tasks

We also broke down the fixation count on different tasks. The box plots of the distributions are shown in Fig. 15 for single attributes and two attributes combined.

Tasks that are not visible on the first page, were fixated significantly less often ($p = 3.73 \times 10^{-8}$, $d = 0.31$, $\Delta = -21.05\%$). Also, tasks not mentioned in a question were fixated significantly less often ($p < 2.0 \times 10^{-16}$, $d = 0.28$, $\Delta = -20.26\%$).

Table 11 Descriptive statistics and hypothesis test results for subsets of tasks (fixation count)

Task attr.	Min.	Med.	Mean	Max.	p	d	Δ
Initially visible	0.00	6.00	7.20	61.00	3.73×10^{-8}	0.31	−1.52
Initially not visible	0.00	5.00	5.68	26.00	(***)	(+)	−21.05%
In question	0.00	7.00	7.59	61.00	$< 2. \times 10^{-16}$	0.28	−1.54
Not in question	0.00	5.00	6.05	42.00	(***)	(+)	−20.26%
Not in quest./init. vis.	0.00	6.00	6.78	27.00	1.31×10^{-8}	0.49	−1.75
Not in quest./not init. vis.	0.00	4.00	5.03	26.00	(***)	(+)	−25.86%
In quest./not init. vis.	0.00	6.00	6.78	19.00	5.86×10^{-6}	0.49	−1.75
Not in quest./not init. vis.	0.00	4.00	5.03	26.00	(***)	(+)	−25.81%
In quest./init. vis.	0.00	7.00	7.76	61.00	4.21×10^{-13}	0.76	−2.74
Not in quest./not init. vis.	0.00	4.00	5.03	26.00	(***)	(++)	−35.24%
In quest./not init. vis.	0.00	6.00	6.78	19.00	0.776	0.00	0.00
Not in quest./init. vis.	0.00	6.00	6.78	27.00			0.07%
In quest./init. vis.	0.00	7.00	7.76	61.00	0.010	0.23	−0.98
Not in quest./init. vis.	0.00	6.00	6.78	27.00	(*)	(+)	−12.65%
In quest./init. vis.	0.00	7.00	7.76	61.00	0.130	0.26	−0.99
In quest./not init. vis.	0.00	6.00	6.78	19.00		(+)	−12.71%

For analyzing the different subsets of tasks clustered by two attributes, we again first used the non-parametric Kruskal–Wallis hypothesis test in order to determine whether the subsets are differing in their means. The test indicated that the means are not equal for the fixation count ($p = 4.515\,709 \times 10^{-12}$). Consequently, we conducted pair-wise unpaired Wilcoxon hypothesis tests for comparing the tasks with different attributes with each other. The results are summarized in Table 11.

Nearly all pairs are differing significantly in their means. Only for the same combinations of attributes which were also not significant with regard to the dwell time are also not significant with the fixation count metric: A task mentioned in a question and placed on the first screen or not the fixation count is not significantly different ($p = 0.130$), although it shows a small effect ($d = 0.26$, $\Delta = -12.71\%$). Also tasks initially visible but not mentioned in a question compared to tasks mentioned in a question but not initially visible showed no significant difference ($p = 0.776$, $d = 0.00$, $\Delta = 0.07\%$).

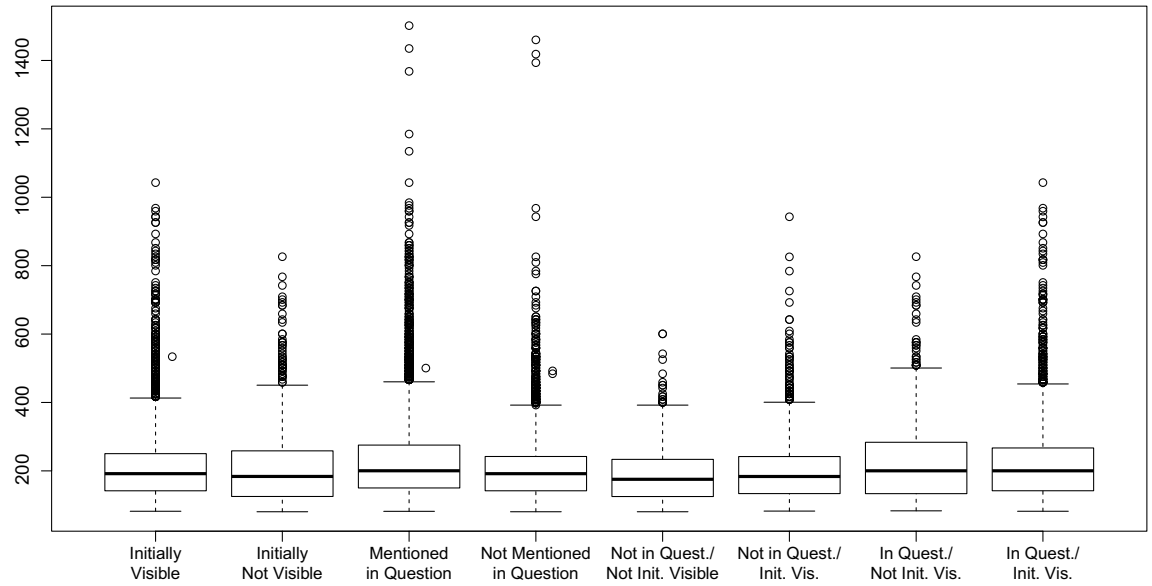

Fig. 16 Fixation duration (ms) for BPMN tasks by all attributes

Table 12 Descriptive statistics and hypothesis test results for subsets of tasks (fixation duration in ms)

Task attr.	Min.	Med.	Mean	Max.	p	d	Δ
Initially visible	81.92	191.86	209.97	1042.86	0.019	0.04	−3.80
Initially not visible	80.48	183.70	206.17	826.27	(*)		−1.81%
In question	81.70	200.32	228.83	1501.92	$< 2. \times 10^{-16}$	0.22	−26.45
Not in question	80.48	191.78	202.38	1460.13	(***)	(+)	−11.56%
Not in quest./init. vis.	82.42	183.57	197.66	943.08	7.13×10^{-3}	0.09	−7.88
Not in quest./not init. vis.	80.48	175.38	189.78	600.73	(**)		−3.99%
In quest./not init. vis.	83.17	200.21	222.32	826.27	1.29×10^{-6}	0.38	−32.54
Not in quest./not init. vis.	80.48	175.38	189.78	600.73	(***)	(+)	−14.64%
In quest./init. vis.	81.92	200.22	221.83	1042.86	4.48×10^{-12}	0.38	−32.05
Not in quest./not init. vis.	80.48	175.38	189.78	600.73	(***)	(+)	−14.45%
In quest./not init. vis.	83.17	200.21	222.32	826.27	2.53×10^{-4}	0.29	−24.66
Not in quest./init. vis.	82.42	183.57	197.66	943.08	(***)	(+)	−11.09%
In quest./init. vis.	81.92	200.22	221.83	1042.86	3.90×10^{-12}	0.28	−24.17
Not in quest./init. vis.	82.42	183.57	197.66	943.08	(***)	(+)	−10.89%
In quest./init. vis.	81.92	200.22	221.83	1042.86	0.584	0.00	0.50
In quest./not init. vis.	83.17	200.21	222.32	826.27			0.22%

5.3.9 Fixation duration of tasks

Subsequently, we analyzed the fixation duration of the tasks. The box plots of the different distributions are shown in Fig. 16 for single attributes and two attributes combined.

When looking at the fixation duration (see Table 12, similar findings can be made. However, in contrast to the dwell time and fixation count, the differences in the mean values for tasks, which are initially visible but not mentioned in a question compared to those that are not initially visible but mentioned in a question, show a significant difference ($p = 2.53e{-}4$, $d = 0.29$, $\Delta = -11.09\%$).

The fixation duration is longer for initially visible tasks compared to those that are not on the first page. The same holds for tasks that are mentioned in a question compared to those that are not asked for.

Analyzing the four additional subsets also yields significant differences in the subset (Kruskal's $p = 1.095\,215 \times 10^{-16}$) and we found significant differences for all remaining combinations. There are two activity type pairs with similar large differences: (i) Tasks that are both not initially visible but differ in whether they are being asked about and (ii)tasks that are initially visible and mentioned in a question compared to tasks that are not initially visible and

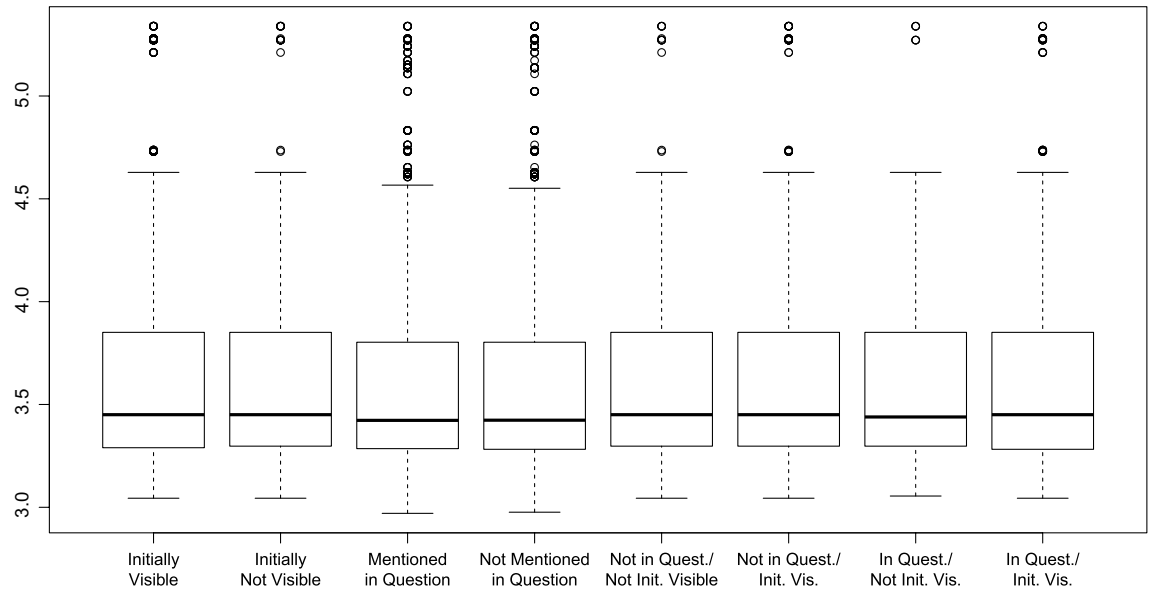

Fig. 17 Pupil diameter (mm) for BPMN tasks by all attributes

Table 13 Descriptive statistics and hypothesis test results for subsets of tasks (pupil diameter in mm)

Task attr.	Min.	Med.	Mean	Max.	p	d	Δ
Initially visible	3.04	3.45	3.64	5.34	0.804	0.01	− 0.01
Initially not visible	3.04	3.45	3.63	5.34			−0.22%
In question	2.97	3.42	3.62	5.34	0.802	0.01	0.00
Not in question	2.98	3.42	3.62	5.34			−0.09%
Not in quest./init. vis.	3.04	3.45	3.63	5.34	0.991	0.00	0.00
Not in quest./not init. vis.	3.04	3.45	3.63	5.34			−0.05%
In quest./not init. vis.	3.05	3.44	3.63	5.34	0.889	0.01	0.01
Not in quest./not init. vis.	3.04	3.45	3.63	5.34			0.19%
In quest./init. vis.	3.04	3.45	3.64	5.34	0.795	0.02	− 0.01
Not in quest./not init. vis.	3.04	3.45	3.63	5.34			−0.29%
In quest./not init. vis.	3.05	3.44	3.63	5.34	0.867	0.02	0.01
Not in quest./init. vis.	3.04	3.45	3.63	5.34			0.24%
In quest./init. vis.	3.04	3.45	3.64	5.34	0.753	0.02	− 0.01
Not in quest./init. vis.	3.04	3.45	3.63	5.34			−0.24%
In quest./init. vis.	3.04	3.45	3.64	5.34	0.711	0.03	− 0.02
In quest./not init. vis.	3.05	3.44	3.63	5.34			−0.47%

not mentioned in a question have very significant p values, a small effect of $d = 0.38$, and a difference around $\Delta\ -\ 14\%$).

5.3.10 Pupil diameter of tasks

The last metric to be broken down is the pupil diameter (see Fig. 17). As before with the dwell time, we used the non-parametric Kruskal-Wallis hypothesis test in order to determine whether the partitions are differing in their means. The test indicated that the means within the partitions are not differing significantly ($p = 0.980\ 8799$).

This metric is completely different than all other metrics. No comparison between different task types yields any significant finding or an effect larger than $d > 0.03$ (see Table 13). The pupil diameter is not significantly affected by task attributes at all.

5.3.11 Subjective preference

At the end of the experiment, subjects were asked for their subjective preference for certain layout options:

- 12 people (57.1%) preferred a horizontal layout and 9 people (42.9%) preferred a vertical layout. A hypothesis test against a null hypothesis that both layouts are equally preferred was conducted by using the z-test and yielded a p value of $p = 0.5127$, which is not significant. Thus, the null hypothesis cannot be rejected.
- When being asked how annoying scrolling (regardless of the layout direction) is, 8 people (38.1%) responded that they perceive scrolling as "very annoying", 8 (38.1%) responded that they perceive scrolling as "annoying" and 5 people (23.8%) responded that they perceive scrolling as "a little annoying". No one answered "not annoying" or "not annoying at all".

6 Interpretation

After presenting the plain experiment results, gathered data and statistics thereof have been presented in the section above, this section will relate these result with the research questions and give an interpretation.

6.1 Evaluation of results and implications

Within this section we will walk through the set of research questions and will provide an interpretation of the measured data in the context of each question group by the two sub-goals.

6.1.1 Diagram layout

This section gives our interpretation of the research question with regard to the overall diagram layout. The findings are summarized in Table 14: For every search question (column 1) we show our hypothesized better and worse layouts (column 2 and 3) and show a summary of our small-grained operational hypothesis. Those, for which the null-hypothesis (no difference) can be rejected are shown in column 4 "Rejected H_0", and those for which the null-hypothesis cannot be rejected are shown in column 5 "Not Rejected H_0". Some operational null-hypothesis are rejected in the expected direction, e.g. the time to answer the questions is indeed significantly shorter. However,

Table 14 Summary of results for research questions concerning diagram layout ((*) results are significantly different but contrary to our hypothesis concerning which layout is better)

	Hypothesized layout		Operational hypothesis	
	Better	Worse	Rejected H_0	Not rejected H_0
RQ1	Horizontal	Vertical	Fixation count H_0^{RQ1-FC} Pupil diameter H_0^{RQ1-PD}	Time to answer, error rate, task efficiency, fixation duration, subjective preference
RQ2	Horizontal	Vertical/scr.	Pupil diameter H_0^{RQ2-PD}	Time to answer, error rate, task efficiency, fixation count, fixation duration
RQ3	Vertical	Horizontal/scr.	Task efficiency(*) H_0^{RQ3-TE} Pupil diameter H_0^{RQ3-PD}	Time to answer, error rate, fixation count, fixation duration
RQ4	Horizontal/snake	Horizontal/scr.	Fixation duration(*) H_0^{RQ4-FD} Pupil diameter H_0^{RQ4-PD}	Time to answer, error rate, task efficiency, fixation count
RQ5	Horizontal/multi-line	Horizontal/scr.	Time to answer H_0^{RQ5-AT} Fixation count H_0^{RQ5-FC} Fixation duration(*) H_0^{RQ5-FD}	Error rate, task efficiency, pupil diameter

there are three cases for which we can reject the null-hypothesis but not in the expected direction. For example, task effectiveness in RQ3 is better for the hypothesized worse layout. Those hypothesis are marked with an asterisk (*). Every research question and the experiment results related to it are explained in detail below.

RQ1: Is "horizontal" or "vertical" layout of BPMN diagrams better understandable?

The first research question is concerned with the understandability of plain horizontal and vertical layouts. The corresponding diagrams in this experiment were D1 (horizontal layout) and D2 (vertical layout). The difference in the means of time required to complete all questions is small $d = 0.28$, $\Delta = 10.03\%$ but not significant, there is no significant difference in the error rate nor the effectiveness. Also, there was no subjective preference for either layout. Therefore, we cannot reject these null hypotheses.

However, the metrics for cognitive load are better for the horizontal layout. The fixation count has a small effect ($d = 0.38$, $\Delta = 17,89\%$) on a significant level ($p = 0.03482$). Also the pupil diameter has a significant difference ($p = 0.03506$) in favor of the horizontal layout. The fixation duration yields a small effect ($d = 0.27$, $\Delta = 4.82\%$), but is not significant.

Therefore, we conclude that the horizontal layout—although not faster to understand—is less demanding to understand as it is imposing a smaller cognitive load.

RQ2: Is "horizontal" or "vertical layout with scrolling" better understandable?

RQ2 is concerned with the understandability of horizontal and vertical layouts when the latter requires scrolling. The corresponding diagrams in this experiment were D3 and D4 respectively.

Again, there was no significant difference for the time required to answer the questions, the error rate nor the task effectiveness. In addition, only one out of the three metrics related to cognitive load showed a significant finding. While both fixation count and fixation duration showed a small effect but not a significant finding, the pupil diameter had a very significant finding with a minor effect size ($p = 0.001374$, $d = 0.17$).

These differences are small to non-existent. As such, we cannot reject the null hypothesis that the horizontal layout without scrolling is as easy to understand both in terms of speed as well as cognitive load.

RQ3: Is "vertical" or "horizontal layout with scrolling" layout better understandable?

The third research question is the reverse of the second one: This time the vertical layout is not being scrolled while the horizontal one is (diagrams D5 and D6). The metrics show a very similar image, though. The time required to complete the questions, the error rate and the task effectiveness have no significant difference. The same is true for the fixation duration. The fixation count shows a small effect size ($d = 0.44$, $\Delta = 12.43\%$) but the difference is not significant.

The pupil diameter has a significant difference ($p = 1.335 \times 10^{-5}$) and a small effect size ($d = 0.18$) as well as the task efficiency ($p = 0.014, d = 0.44$). These differences are higher than with RQ2 but however small. As such, we can only find (too) weak arguments for rejecting the overall hypothesis.

RQ4: Is a horizontal snake layout better understandable than a diagram that requires scrolling?

When comparing a larger, horizontally laid out diagram with a snake layout in comparison to a horizontal layout

that requires scrolling (diagrams D7 and D8), the differences are larger than with the previous diagrams.

The answer time is significantly shorter ($p = 0.0484$, $d = 0.3$). The metrics related to cognitive load however are mixed. Fixation count yields an effectless and not significant difference in favor of the snake layout. Pupil diameter has the same direction but indicates a small effect with a significant finding. In contrast, the fixation duration is shorter for the scrollable layout with a very significant finding and medium effect size ($p = 0.001859$, $d = 0.65$).

With these measurements and test results, we reject the null hypothesis and conclude that the snake layout is faster to understand but we cannot reject the null hypothesis that the snake layout is as demanding as the scrollable horizontal layout because for the latter the measurements are inconclusive. In contrast, they point to more cognitive effort required to understand the diagram (although understand it faster).

RQ5: Is a horizontal multi-line layout better understandable than a diagram that requires scrolling?

When comparing a large, horizontally laid out diagram arranged as a multi-line layout compared to a large horizontal layout that requires scrolling (diagrams D9 and D10), the findings are similar to those of the snake layout.

The answer time is very significantly different ($p = 0.0049$) and shows a small effect ($d = 0.39$) in favor of the multi-line layout. Error rate and task efficiency show no significant difference.

Cognitive load metrics are undecided. Fixation count has a non-significantly better ($p = 0.05583$) small effect ($d = 0.26$) for the multi-line layout, pupil diameter has no effect ($d = 0.01$) and the difference is not significant ($p = 0.4948$). Fixation duration is very significantly better for the horizontal layout that requires scrolling and yields a medium effect ($d = 0.73$, $\Delta = -10.43\%$).

With these measurements and test results, we reject the null hypothesis that conclude that the multi-line layout is faster to understand but we cannot reject the null hypothesis that the snake layout is as cognitively demanding as the scrollable horizontal layout because for the latter the measurements are inconclusive.

6.1.2 Task attributes

This sections presents our interpretation of the experiment results that are concerned with the sub-goal of task attributes. The next two research questions are not concerned with the overall diagram layout but with specific positions of tasks.

RQ6: Are tasks that are located outside of the initially visible area read less?

The sixth research question was concerned with the tasks that are initially visible or not, i.e., whether the model reader was required to scroll to a specific tasks: Tasks that were positioned on the first page were looked at more often. Dwell time, i.e., the combined time spent looking at a task, was highly significantly larger ($p = 4.86 \times 10^{-6}$) with a small effect size ($d = 0.27$) and a relative mean difference of $\Delta = 21.63\%$. Also the fixation count ($p = 0.0001461$, $d = 0.23$, $\Delta = 17.56\%$) and the fixation duration ($p = 5.044 \times 10^{-7}$, $d = 0.10$, $\Delta = 5.03\%$) showed a significant difference between these tasks. Only the pupil diameter had no significant difference and no effect. Thus, all operational null hypothesis but the one concerned with the pupil diameter can be rejected (H_0^{RQ6-DT}, H_0^{RQ6-FC}, H_0^{RQ6-FD}, H_0^{RQ6-PD}).

Therefore, our overall null hypothesis is rejected and we conclude that tasks outside the initially visible area are read less based on the longer dwell time and the higher fixation count. In addition, model readers seem to invest less mental effort in tasks outside the initially visible area indicated by a smaller average fixation duration.

RQ7: Are tasks read selectively based on the job of the model reader?

When grouping tasks based on whether they are mentioned in a question regarding the diagram or not, a similar picture like the tasks on the initial screen emerges: The dwell time is significantly longer for tasks mentioned in questions ($p = 1.701 \times 10^{-23}$, $d = 0.27$, $\Delta = 26.70\%$), fixation count is significantly larger ($p = 1.625e-15$, $d = 0.28$, $\Delta = 20.68\%$), and the fixation duration is longer ($p = 5.594 \times 10^{-32}$, $d = 0.21$, $\Delta = 11.30\%$) for tasks that are mentioned in a question. Like in RQ6 no difference with regards to the pupil diameter was found. Thus, all operational null hypothesis but the one concerned with the pupil diameter can be rejected (H_0^{RQ7-DT}, H_0^{RQ7-FC}, H_0^{RQ7-FD}, H_0^{RQ7-PD}).

Thus, we reject the null hypothesis and conclude that tasks that are mentioned in a question, are read more often and longer, and tasks that are not mentioned in a question are read less. Also, more mental effort is invested in tasks mentioned in a question as indicated by the increased fixation duration.

When diving deeper into the difference concerning tasks with different attributes, we found significant differences in dwell time, fixation duration, and fixation count for nearly every combination of *initially visible* or not and *mentioned in question* or not. The general rule is that tasks that are on the first screen and/or mentioned in a question are read more often (fixation count) and longer (dwell time) and with more intensity (fixation duration) than those that do not share these attributes. The largest difference is between tasks that are mentioned in a question and are visible on the first screen compared to tasks that are not mentioned in a question and are not on the first screen: In this case, the maximal difference in all metrics was measured (Dwell

time: $p < 2.0e{-}16$, $d = 0.95$, $\Delta = -42.69\%$, fixation count: $p = 6.75e{-}14$, $d = 0.76$, $\Delta = -35.13\%$, fixation duration: $p < 2.0e{-}16$, $d = 0.46$, $\Delta = -17.05\%$). This shows that these two attributes combined are very powerful in directing the reader's attention.

There is only one combination of task attributes that does not show significant differences in the dwell time and fixation count metrics: If an task is mentioned in a question, there is no significant difference whether it is on the first page or not. This means that concerning the reader's attention the task (task being mentioned in the question) supersedes the position on the first screen.

6.2 Consequences for BPMN modeling

Based on the answers to our research questions, some modeling guidelines can be derived, which help BPMN modelers to design process models that are better and/or quicker to understand. The following rules can complement any modeling guidelines in use:

Use a horizontal layout Although there is only a negligible difference in the time required to understand a BPMN model based on the layout direction, cognitive load metrics were considerably better for the horizontal layout. As such, BPMN models should be arranged horizontally.
Maximize number of visible elements by using Snake or Multiline Layouts If a horizontal layout gets too large to fit on a single screen, break the layout visually in a snake or multi-line layout. Although the mental effort for understanding such a model is not affected, the time required to understand the model is significantly reduced.
Put important elements on first page If a model does not fit on a single screen or page (e.g., because BPMN models are auto-formatted or are printed on paper), consider that readers are focusing tasks on the first page most. Thus, a modeler should place important tasks on the first page.
Consider that model readers focus heavily on elements of interest Consider that model readers are searching for information they require (at least in our task-based experiment). If you think that a piece of information is important but not necessarily in-scope of a typical model reader, guide the reader more prominently to this information.

6.3 Limitations of study

Like every other empirical inquiry, the presented experiment, its results and interpretation are subject to threats of validity. Consequently, possible threats are discussed along the classification of Cook and Campbell [10] in this section.

6.3.1 Conclusion validity

Since the data does not follow a normal distribution, we used the non-parametric Wilcoxon test. Although it is not as powerful as parametric tests such as the t-test, it does not have any assumptions on the data that could be violated. We only used paired tests when appropriate, i.e., when comparing data from the same subject on diagrams with different layouts. Nonetheless, the sample size of 21 participants limits the statistical power of our results.

By using the GQM approach and defining our goals, research questions and metrics before conducting the experiment, we prevented risking fishing for specific results. To improve reliability of our results, we used eye tracking as an objective, automated, easily reproducible way of measuring metrics without any subjective bias. However, eye tracking data is subject to errors due to individual differences between subjects, changes in lighting or positioning in front of the laptop. We also encountered subjects that could not be tracked properly with the eye tracker that we used in our experiment. Other factors that can influence the results such as task experience are difficult to control. To observe this possible effect, we asked participants to rate their experience with BPMN and UML Activity Diagrams. The ratings show that participants mostly had an average prior knowledge of BPMN and UML Activity Diagrams. Thus, prior knowledge should only have a minor effect on the results.

There is no threat to the reliability of treatment implementation. By automatically guiding participants through the experiment with a small software application, showing them the experiment contents such as the task introduction and diagrams in the exact same way, we provided a fully standardized setup. No manual treatment was given. Additionally, the experiment took place in the same room for all subjects without any noise or disruptions.

6.3.2 Internal validity

By using a within-group design in our experiment we avoided threats caused by interactions with selection because every participant was confronted with all considered layout styles. Our experiment was not subject to any learning effects since subjects could not take any knowledge from answering one question to the next one. Besides, they were not given any feedback on their test results. Compared diagrams directly followed one another. Therefore participants had comparable experience when answering the respective questions.

To ensure the quality and understandability of all diagrams and questions that were used in the experiment, we performed a dry run with three professionals before the actual execution. That way potential ambiguities or misunderstandings could be resolved beforehand.

However, there might be a small bias because the subjects volunteered to participate in the experiment. Volunteers are usually more motivated than the whole population. Since the participants were randomly assigned to the two groups and there was no control and treatment group, this does not have a major effect on the results. We counteracted ambiguity about direction of causal influence by always comparing the exact same diagrams (apart from task naming) and only changing one controlled variable, i.e., the layout.

Diffusion of imitation of treatments, compensatory equalization of treatments, compensatory rivalry and resentful demoralization, as well as statistical regression do not apply to our experiment because we did not have a control group and did not classify the subjects based on previous studies. Three participants' data could not be used because of recording issues with the eye tracker. Since ability to being eye tracked is not a factor in diagram understanding, this is no threat to validity.

6.3.3 Construct validity

All layout options were clearly defined and the tested diagrams created accordingly. In order to prevent biases caused by a specific process context and to minimize the time spent on the tasks, we chose to use single letters as labels to keep the diagrams simple and easy to understand. Hence, the results might differ for larger, more complex models that are more textual. Because we only used one diagram for each layout variant, we have a mono-operation bias. This decision was made to be able to validly compare the respective layouts.

To improve construct validity and counteract a mono-method bias, we measured many different metrics and built our results mostly on objective eye tracking measures, instead of only taking subjective data into account. None of the subjects knew about the research questions or hypotheses prior to the experiment. All the subjective ratings were given in a questionnaire at the very end of the experiment, so that they did not influence any of the quantitative metrics on the previously shown diagrams. However, since the diagrams only differed in layout and size, subjects might have been able to guess the study objective and thus the subjective preference could bias the objective metrics.

We counteracted restricted generalizability across constructs by measuring different aspects of layout quality, i.e. speed, cognitive load and subjective preference. That way we could control whether the improvement of one attribute might have a negative effect on another and vice versa.

To prevent a bias caused by certain prior expectations by the experimenter, the whole experiment was guided by the application software and without human interactions. The only step where the experimenter interacted with subjects was during the calibration of the eye tracker. Particularly, all contents such as introduction, consent, questions and diagrams were displayed on screen. Therefore, they were the same for all participants.

6.3.4 External validity

As with any experiment, this experiment had to establish a controlled environment which differs from real-life projects. These differences may or may not influence the generalizability of the results.

By selecting industry developers as subjects, we produced a good level of realism. Nonetheless, the results might not be generalizable for other roles that are also concerned with modeling and reading BPMN diagrams, especially non-technical roles.

The use of small task labels with only one letter was chosen in order to isolate the effect of layout characteristics. However, such diagrams do not resemble real diagrams. Hence, the influence of label phrasing and label length might be larger than the effects found in this experiment, which poses a threat to generalizability. Further research should be done on larger, more realistic BPMN diagrams including textual labels. Besides, tasks that were displayed on the edge of the screen, i.e., where the label was still readable without scrolling, but parts were already initially hidden, were also considered as initially visible. This might have an effect on the mean dwell times for tasks that did not require scrolling. This was only the case for one task in two diagram pairs. The effect, therefore, is rather small.

The BPMN diagrams are compliant to the most basic BPMN Level 1. Thus, they do not include more advanced elements like boundary events or pools and lanes. These elements might influence the layout and aesthetics of different layout directions. As such, our results might not be generalizable to BPMN diagrams using more advanced BPMN constructs.

Also the BPMN diagrams contained only a limited number of tasks and gateways. Especially with execution projects larger diagrams might be more common, although there are also modeling guidelines, which aim at restricting the size of a diagram in terms of visible elements; for example, the eCH modeling guidelines [5, p. 11] impose a "9–15" activity limit. Mendling et al. have identified thresholds for complexity metrics, beyond which modeling errors are more likely [35]. The same is likely be true for reading models, i.e., after a certain model complexity is reached, comprehensibility suffers. Our diagrams have probably not teared such thresholds because our error rate is very low and nearly all questions were answered correctly. Thus, our findings might not be generalizable to larger and/or more complex diagrams.

All diagrams were created with current tools and the display on screen is representative for what is used in practice even if other possibilities are also common.

6.4 Inferences

Based on answers obtained to our hypotheses we infer that for reading BPMN diagrams on a screen, horizontal layout is better comprehensible than a vertical layout, and diagrams which require scrolling are less comprehensible than those which do not. Moreover, we can infer that it is better to place all elements on one screen with a more complex layout (e.g., snake or multi-line layouts) than require the model reader to scroll. This is based on our findings that both more complex layouts are being better understandable than horizontal scrolling counterparts.

However, we do not generalize this to other media, especially tablets, which offer other means for scrolling and different orientation of the visible area (computer screens have a horizontal orientation, while paper and tablets can be turned around easily); Turetken et al. have already shown that understandability of paper was superior to on-screen presentation in their experiment [59, 60].

We infer that our findings generalize to people familiar with "box and arrows" diagrams for denoting processes and other control-flows. We do not generalize this to people without such experience. More research into how untrained people perceive these diagrams is required to either generalize or explicitly demarcate the background of model readers.

Also supported by the findings of Petrusel et al. [42] concerning task-based, relevant areas, we generalize that elements being important for the task at hand—in our case because they are mentioned in a question—are being looked at more often and more frequently. Furthermore, we see no reason why our findings with regard to the placement on the first page or on a following page should not generalize to other diagrams, model readers or models. This is also consistent with results from the Web design domain. Fessenden [16] found that Web users spend 57% of their time on the initially visible screen area.

However, we expect differences with other modeling notations. Jost et al. [24] have already compared BPMN, UML Activity Diagrams and Event-driven Process Chains for business processes of different complexity. They found that depending on the complexity different notations were more comprehensible than others. Due to prescribed modeling directions (Event-driven Process Chains are standardized to be laid out vertically) and different visual syntax, especially sizes of their symbols and containers, results and valid options may vary for different notations.

6.5 Lessons learned

Like with every activity, one gains experience. In order to help other researchers avoid pitfalls, which we encountered during this experiment, we share our experiences in this section.

- Always verify data correctness and validity manually (at least for random data samples) before analyzing the results and performing statistical tests. Statistical tools provide nice visualizations of the data and some likely errors can be spotted there. However, the data—especially when the measurements are potentially imprecise like eye tracking data—should be validated first. Although eye movement tracking was closely supervised by the experimenter during the conduction, we found data that needed to be excluded or manually corrected afterwards.
- Plan ahead the data analysis, concrete research questions and metrics before designing the experiment material and before executing the study. The use of GQM or other approaches *before* setting up the experiment is a great help. With the next experiment we would try to go a step further and script the whole data analysis process beforehand and test it with either dummy data or in a pre-experiment.
- Plan for removing subjects when using eye tracking. When conducting eye tracking experiments, it is very likely to encounter subjects that cannot be properly tracked. Thus, the initial sample size needs to be large enough to compensate such cases. This is a problem because of the effort required by eye tracking experiments. Furthermore, experiment execution cannot be parallelized (if only one eye tracker is available), which adds to the required time for such research projects.
- Utilize your research team. Validate and review all materials and scripts. The more persons inspect the experiment material, the more mistakes are spotted and the more discussion is fostered, which in turn leads to new ideas.
- Use the eye tracker as an incentive for tech-savvy participants. At first we hoped that we will find enough participants for our experiment. However, people were interested in seeing an eye tracker in action that we had not enough time for giving every volunteer a time-slot.
- Metrics indirectly measuring cognitive load may conflict with each other. When looking at fixation duration, fixation count and pupil diameter, the different metrics vary in their results and consequently possible interpretations. In general, we found the pupil diameter only applicable to whole diagrams while fixation count and fixation duration also gave differences when looking at diagram element level. We suppose that the pupil diameter does not change as quickly when hovering over different diagram elements as required to spot differences in cognitive load.
- While Pupil Diameter is a metric that has been used as an indirect indicator of cognitive load. However, it seems to be a metric that cannot indicate quick changes. While all

other metrics for different tasks showed significant differences, pupil diameter remained the same across all task combinations. Perhaps the pupil diameter changes slowly with the cognitive load and not as quick as the model reader jumps between different objects in a diagram.

7 Conclusions and outlook

7.1 Possible future research

Based on the findings and open questions of the presented experiment, there are possible follow ups for future research: Regarding the layout, it is interesting to analyze whether snake or multi-line layouts are better comprehensible. This could not be answered by our experiment design and is left for future research. Similar studies can also be made for other layout variants, e.g., large vertical diagrams that require scrolling or are laid out using snake or multi-line layouts.

Current research work focuses on simple BPMN diagrams, which more or less are comprised of BPMN Level 1 elements. Including more complex branching mechanisms in the layout of eye tracking studies, e.g., BPMN's boundary events, would extend the current state of knowledge. This can be further extended by using more realistic labels in order to quantify how large the effect of labels is in contrast to layout orientation.

Another angle of differentiation would be the impact of different layouts on different devices and media. Exploratory analysis has shown that participants scrolled horizontally by using the scroll bar and vertically by using the mouse wheel. When using tablets, scrolling to both directions would be the done by the same mechanism. The impact of this is currently unknown. Only a study comparing on-screen and paper process models was conducted, which found that paper is better comprehensible [60].

Our findings can influence the field of automatic creation of layouts for BPMN process diagrams as well. With the findings of this experiment, the problem of how to honor different page/screen sizes in automatic layouts emerges. Such algorithms would need to switch to a snake or multi-line layout if this allows to layout the whole diagram on one page. This in turn means that layout algorithms need to be aware of the available space on the target medium.

Replication studies with a more business-oriented population can strengthen the generalizability of the results. Furthermore, replication and thus a larger combined sample size can also increase the confidence in the results.

There are other measures available to reduce the space required by BPMN diagrams. For example, elements and fonts can be made smaller. Experiments with different "zoom levels" of diagrams could shed light on the usefulness of such modeling measures.

Because Mandarin-speaking people relate time more with the y-axis [6], replication of this experiment with a Mandarin-speaking population can show whether this impacts the findings or not—thereby either strengthening the generalizability or establishing boundaries for the results.

7.2 Conclusions

This article presented the results of an eye tracking experiment for gaining insight into the effects of different BPMN diagram layouts on understandability. The experiment was conducted during a training event of a company and utilized eye tracking data from 21 professional software developers.

Significant differences were found in the required cognitive load, for which a horizontal layout scored better than a vertical layout. If a BPMN model becomes too large, more complex layouts—like snake and multi-line layouts—reduce the time required for giving answers to a model compared to a large model that requires extensive horizontal scrolling. Furthermore, BPMN elements are read more intensively if they are necessary for the task at hand and they are read more often if they are placed on the first screen.

Thus, models should be laid out left-to-right if possible—which also makes most use of computer screen estate for most BPMN diagrams—and switch to a snake or multi-line layout if the process models get too large to fit on one screen in order to utilize the whole screen until moving to a scrollable diagram. However, with diagrams that are comparatively small we surprisingly found no differences in the understandability of diagrams when they need to be scrolled in one direction but would not require scrolling when laid out differently.

We also found inconsistencies in the cognitive load metrics. Those metrics did not result in consistent findings. For example, pupil diameter seems to be a better indicator for a whole diagram than for single diagram elements.

This work adds further knowledge of empirical inquiries to the existing body of knowledge concerned with BPMN layout and its implications for comprehensibility.

Operational hypothesis

RQ1: Is a "horizontal" or a "vertical" layout (both without scrolling) of BPMN diagrams better understandable?

H_0^{RQ1-AT}: The answer time for questions is the same.

H_1^{RQ1-AT}: The answer time for questions is shorter for the horizontal layout.
H_0^{RQ1-E}: The number of errors made while answering the questions is the same.
H_1^{RQ1-E}: The number of errors made while answering the questions is lower for the horizontal layout.
H_0^{RQ1-TE}: The task efficiency of both layouts is the same.
H_1^{RQ1-TE}: The task efficiency for the horizontal layout is higher.
H_0^{RQ1-P}: The subjective layout preference is the same.
H_1^{RQ1-P}: The horizontal layout is preferred better.
H_0^{RQ1-FC}: The fixation count is the same on both layouts.
H_1^{RQ1-FC}: The fixation count for the horizontal layout is less.
H_0^{RQ1-FD}: The fixation duration is the same on both layouts.
H_1^{RQ1-FD}: The fixation duration for the horizontal layout is shorter.
H_0^{RQ1-PD}: The pupil diameter is the same on both layouts.
H_1^{RQ1-PD}: The pupil diameter for the horizontal layout is smaller.

RQ2: Is a "horizontal" or a "vertical layout with scrolling" better understandable?

H_0^{RQ2-AT}: The answer time for questions is the same.
H_1^{RQ2-AT}: The answer time for questions for the horizontal layout is shorter.
H_0^{RQ2-E}: The number of errors made while answering the questions is the same.
H_1^{RQ1-E}: The number of errors made while answering the questions is lower for the horizontal layout.
H_0^{RQ2-TE}: The task efficiency for both layouts is the same.
H_1^{RQ1-TE}: The task efficiency for the horizontal layout is higher.
H_0^{RQ2-FC}: The fixation count is the same on both layouts.
H_1^{RQ2-FC}: The fixation count for the horizontal layout is less.
H_0^{RQ2-FD}: The fixation duration is the same on both layouts.
H_1^{RQ2-FD}: The fixation duration for the horizontal layout is shorter.
H_0^{RQ2-PD}: The pupil diameter is the same on both layouts.
H_1^{RQ2-PD}: The pupil diameter for the horizontal layout is smaller.

RQ3: Is a "vertical" or a "horizontal layout with scrolling" better understandable?

H_0^{RQ3-AT}: The answer time for questions is the same.
H_1^{RQ3-AT}: The answer time for questions is shorter for the vertical layout.
H_0^{RQ1-E}: The number of errors made while answering the questions is the same.
H_1^{RQ3-E}: The number of errors made while answering the questions is lower for the vertical layout.
H_0^{RQ1-TE}: The task efficiency for both layouts is the same.
H_1^{RQ3-TE}: The task efficiency for the vertical layout is higher.
H_0^{RQ3-FC}: The fixation count is the same on both layouts.
H_1^{RQ3-FC}: The fixation count for the vertical layout is less.
H_0^{RQ3-FD}: The fixation duration is the same on both layouts.
H_1^{RQ3-FD}: The fixation duration for the vertical layout is shorter.
H_0^{RQ3-PD}: The pupil diameter is the same on both layouts.
H_1^{RQ3-PD}: The pupil diameter for the vertical layout is smaller.

RQ4: Is a horizontal snake layout better understandable than a horizontal layout with scrolling?

H_0^{RQ4-AT}: The answer time for questions is the same.
H_1^{RQ4-AT}: The answer time for questions is shorter for the horizontal layout.
H_0^{RQ4-E}: The number of errors made while answering the questions is the same.
H_1^{RQ1-E}: The number of errors made while answering the questions is lower for the horizontal snake layout.
H_0^{RQ4-TE}: The task efficiency for both layouts is the same.
H_1^{RQ1-TE}: The task efficiency for the horizontal snake layout is higher.
H_0^{RQ4-FC}: The fixation count is the same on both layouts.
H_1^{RQ4-FC}: The fixation count for the horizontal snake layout is less.
H_0^{RQ4-FD}: The fixation duration is the same on both layouts.
H_1^{RQ4-FD}: The fixation duration for the horizontal snake layout is shorter.
H_0^{RQ4-PD}: The pupil diameter is the same on both layouts.
H_1^{RQ4-PD}: The pupil diameter for the horizontal snake layout is smaller.

RQ5: Is a horizontal multi-line layout better understandable than a horizontal layout with scrolling?

H_0^{RQ5-AT}: The answer time for questions is the same.
H_1^{RQ5-AT}: The answer time for questions is shorter for the horizontal multi-line layout.

H_0^{RQ5-E}: The number of errors made while answering the questions is the same.
H_1^{RQ1-E}: The number of errors made while answering the questions is lower for the horizontal multi-line layout.
H_0^{RQ5-TE}: The task efficiency for both layouts is the same.
H_1^{RQ1-TE}: The task efficiency for the horizontal multi-line layout is higher.
H_0^{RQ5-FC}: The fixation count is the same on both layouts.
H_1^{RQ5-FC}: The fixation count for the horizontal multi-line layout is less.
H_0^{RQ5-FD}: The fixation duration is the same on both layouts.
H_1^{RQ5-FD}: The fixation duration for the horizontal multi-line layout is shorter.
H_0^{RQ5-PD}: The pupil diameter is the same on both layouts.
H_1^{RQ5-PD}: The pupil diameter for the horizontal multi-line layout is smaller.

RQ6: Are tasks that are located outside of the initially visible area read less?

H_0^{RQ6-DT}: The dwell time for all questions is the same.
H_1^{RQ6-DT}: The dwell time of initially visible tasks is higher.
H_0^{RQ6-FC}: The fixation count is the same for all tasks.
H_1^{RQ6-FC}: The fixation count of initially visible tasks is higher.
H_0^{RQ6-FD}: The fixation duration is the same for all tasks.
H_1^{RQ6-FD}: The fixation duration of initially visible tasks is higher.
H_0^{RQ6-PD}: The pupil diameter is the same for all tasks.
H_1^{RQ6-PD}: The pupil diameter of initially visible tasks is higher.

RQ7: Are tasks read selectively based on the job of the model reader?

H_0^{RQ7-DT}: The dwell time for all questions is the same.
H_1^{RQ7-DT}: The dwell time of tasks mentioned in questions is higher.
H_0^{RQ7-FC}: The fixation count is the same for all tasks.
H_1^{RQ7-FC}: The fixation count of tasks mentioned in questions is higher.
H_0^{RQ7-FD}: The fixation duration is the same for all tasks.
H_1^{RQ7-FD}: The fixation duration of tasks mentioned in questions is higher.
H_0^{RQ7-PD}: The pupil diameter is the same for all tasks.
H_1^{RQ7-PD}: The pupil diameter of tasks mentioned in questions is higher.

Acknowledgements We thank all participants of this experiment and SMI (Senso-Motoric Instruments) for providing an SMI RED-m eye tracker in order to support our studies. We also thank the reviewers, who did a very good job and thereby helped to improve this paper with valuable and constructive feedback.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

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

References

1. Ahrens M, Schneider K, Kiesling S (2016) How do we read specifications? experiences from an eye tracking study. In: International working conference on requirements engineering: foundation for software quality. Springer, pp 301–317
2. Basili VR (1993) Applying the goal/question/metric paradigm in the experience factory. Software quality assurance and measurement: a worldwide perspective, pp 21–44
3. Berli W, Lübke D, Möckli W (2014) Terravis - large scale business process integration between public and private partners. In: Plödereder E, Grunske L, Schneider E, Ull D (eds) Lecture notes in informatics (LNI), proceedings INFORMATIK 2014, vol P-232. Gesellschaft für Informatik e.V., pp 1075–1090
4. Bernstein V, Soffer P (2015) Advanced information systems engineering workshops: CAiSE 2015 international workshops, Stockholm, Sweden, June 8-9, 2015, Proceedings, chap. How does it look? Exploring meaningful layout features of process models. Springer, Cham, pp 81–86
5. Birchler A, Bosshart E, Märki M, Opitz P, Pauli J, Rigert B, Sandoz Y, Schaffroth M, Spöcker N, Tanner C, Walser K, Widmer T (2014) eCH-0158 BPMN-Modellierungskonventionen für die öffentliche Verwaltung. https://www.ech.ch/dokument/fb5725cb-813f-47dc-8283-c04f9311a5b8
6. Boroditsky L (2008) Do English and Mandarin speakers think differently about time? In: Proceedings of the annual meeting of the cognitive science society, vol 30
7. Brill O, Schneider K, Knauss E (2010) Videos vs. use cases: Can videos capture more requirements under time pressure?. In: International working conference on requirements engineering: foundation for software quality. Springer, pp 30–44
8. Busjahn T, Bednarik R, Schulte C (2014) What influences dwell time during source code reading?: Analysis of element type and frequency as factors. In: Proceedings of the symposium on eye tracking research and applications, ETRA '14. ACM, New York, NY, USA, pp 335–338. https://doi.org/10.1145/2578153.2578211

9. Compeau D, Marcolin B, Kelley H, Higgins C (2012) Research commentary-generalizability of information systems research using student subjects-a reflection on our practices and recommendations for future research. Inf Syst Res 23(4):1093–1109
10. Cook TD, Campbell DT (1979) Quasi-experimentation: design & analysis issues for field settings. Houghton Mifflin Company, Boston
11. Dikici A, Turetken O, Demirors O (2018) Factors influencing the understandability of process models: a systematic literature review. Inf Softw Technol 93:112–129
12. Djamasbi S, Mehta D, Samani A (2012) Eye movements, perceptions, and performance. In: Proceedings of the eighteenth Americas conference on information systems (AMCIS), pp 1–7
13. Djamasbi S, Siegel M, Skorinko J, Tullis T (2011) Online viewing and aesthetic preferences of generation y and baby boomers: testing user website experience through eye tracking. Int J Electron Commer 15(4):121–158
14. Effinger P (2011) Business process model and notation: third international workshop, BPMN 2011, Lucerne, Switzerland, November 21–22, 2011. Proceedings, chap. Layout patterns with BPMN semantics. Springer, Berlin, pp. 130–135
15. Effinger P, Decker G (2010) Graph drawing: 17th international symposium, GD 2009, Chicago, IL, USA, September 22–25, 2009. Revised papers, chap. Layout techniques coupled with Web2.0-based business process modeling. Springer, Berlin, pp 417–418
16. Fessenden T (2018) Scrolling and attention. Nielson Normam Group, on April 15. https://www.nngroup.com/articles/scrolling-and-attention/
17. Figl K (2017) Comprehension of procedural visual business process models. Bus Inf Syst Eng 59(1):41–67
18. Figl K, Strembeck M (2014) On the importance of flow direction in business process models. In: 2014 9th international conference on software engineering and applications (ICSOFT-EA). IEEE, pp 132–136
19. Figl K, Strembeck M (2015) Findings from an experiment on flow direction of business process models. In: EMISA 2015
20. Genon N, Heymans P, Amyot D (2011) Software language engineering: third international conference, SLE 2010, Eindhoven, The Netherlands, October 12–13, 2010, Revised selected papers, chap. Analysing the cognitive effectiveness of the BPMN 2.0 visual notation. Springer, Berlin, pp 377–396
21. Goldberg J, Kotval X (1999) Computer interface evaluation using eye movements: methods and constructs. Int J Ind Ergon 24:631–645. https://doi.org/10.1016/S0169-8141(98)00068-7
22. Haisjackl C, Burattin A, Soffer P, Weber B (2017) Visualization of the evolution of layout metrics for business process models. In: 14th conference in the field of business process management. Springer, pp 449–460
23. Hess EH, Polt JM (1964) Pupil size in relation to mental activity during simple problem-solving. Science 143(3611):1190–1192
24. Jošt G, Huber J, Heričko M, Polančič G (2016) An empirical investigation of intuitive understandability of process diagrams. Comput Stand Interfaces 48:90–111
25. Kitzmann I, König C, Lübke D, Singer L (2009) A simple algorithm for automatic layout of BPMN processes. In: 2009 IEEE conference on commerce and enterprise computing. IEEE, pp 391–398
26. Kretschmann K (2019) Investigating the flow direction in business process models: an eye tracking study. Bachelor's thesis. http://dbis.eprints.uni-ulm.de/1848/
27. Lindland OI, Sindre G, Solvberg A (1994) Understanding quality in conceptual modeling. IEEE Softw 11(2):42–49. https://doi.org/10.1109/52.268955
28. Lübke D (2015) Using metric time lines for identifying architecture shortcomings in process execution architectures. In: 2015 IEEE/ACM 2nd international workshop on software architecture and metrics (SAM). IEEE, pp 55–58
29. Lübke D, van Lessen T (2016) Modeling test cases in BPMN for behavior-driven development. IEEE Softw 2016:17–23
30. Lübke D, Wutke D (2021) Analysis of prevalent BPMN layout choices on github. In: Proceedings of the 13th Central European workshop on services and their composition (ZEUS 2021) (**accepted**)
31. Lübke D, Ahrens M, Schneider K (2021) Dataset and materials for "influence of diagram layout and scrolling on understandability of BPMN processes". https://doi.org/10.5281/zenodo.4557963
32. Mendling J, Recker J, Reijers HA, Leopold H (2019) An empirical review of the connection between model viewer characteristics and the comprehension of conceptual process models. Inf Syst Front 21(5):1111–1135
33. Mendling J, Reijers HA, van der Aalst WM (2010) Seven process modeling guidelines (7PMG). Inf Softw Technol 52(2):127–136
34. Mendling J, Strembeck M, Recker J (2012) Factors of process model comprehension-findings from a series of experiments. Decis Support Syst 53(1):195–206
35. Mendling J, Sánchez-González L, García F, La Rosa M (2012) Thresholds for error probability measures of business process models. J Syst Softw 85(5):1188–1197
36. Moody DL (2009) The physics of notations: toward a scientific basis for constructing visual notations in software engineering. IEEE Trans Softw Eng 35(6):756–779
37. Moody DL (2011) Why a diagram is only sometimes worth a thousand words: an analysis of the BPMN 2.0 visual notation
38. Nick M, Tautz C (1999) Practical evaluation of an organizational memory using the goal-question-metric technique. In: German conference on knowledge-based systems. Springer, pp 138–147
39. Object Management Group (OMG) (2011) Business process model and notation (BPMN), version 2.0. Technical report, Object Management Group (OMG). http://www.omg.org/spec/BPMN/2.0/
40. Ottensooser A, Fekete A, Reijers HA, Mendling J, Menictas C (2012) Making sense of business process descriptions: an experimental comparison of graphical and textual notations. J Syst Softw 85(3):596–606
41. Paas F, Tuovinen JE, Tabbers H, Gerven PWMV (2003) Cognitive load measurement as a means to advance cognitive load theory. Educ Psychol 38(1):63–71
42. Petrusel R, Mendling J (2013) Eye-tracking the factors of process model comprehension tasks. In: Salinesi C, Norrie MC, Pastor Ó (eds) Advanced information systems engineering. Springer, Berlin, pp 224–239
43. Petrusel R, Mendling J, Reijers HA (2016) Task-specific visual cues for improving process model understanding. Inf Softw Technol 79:63–78
44. Poole A, Ball LJ (2006) Eye tracking in HCI and usability research. In: Encyclopedia of human computer interaction. IGI Global, pp 211–219
45. Purchase HC (2002) Metrics for graph drawing aesthetics. J Vis Lang Comput 13(5):501–516
46. Purchase HC, Allder J, Carrington D (2002) Graph layout aesthetics in UML diagrams: user preferences. J Graph Algorithms Appl 6(3):255–279. https://doi.org/10.7155/jgaa.00054
47. Purchase HC, Carrington D, Allder JA (2000) Theory and application of diagrams: first international conference, diagrams 2000 Edinburgh, Scotland, UK, September 1–3, 2000 Proceedings, chap. Experimenting with aesthetics-based graph layout. Springer, Berlin, pp 498–501
48. Purchase HC, Carrington D, Allder JA (2002) Empirical evaluation of aesthetics-based graph layout. Empir Softw Eng 7(3):233–255

49. Recker J (2013) Empirical investigation of the usefulness of gateway constructs in process models. Eur J Inf Syst 22(6):673–689
50. Scholz T, Lübke D (2019) Improving automatic BPMN layouting by experimentally evaluating user preferences. In: Rocha Á, Adeli H, Reis LP, Costanzo S (eds) New knowledge in information systems and technologies. Springer, Cham, pp 748–757
51. Schrepfer M, Wolf J, Mendling J, Reijers HA (2009) The impact of secondary notation on process model understanding. In: The practice of enterprise modeling. Springer, pp 161–175
52. Sharafi Z, Shaffer T, Sharif B, Gueheneuc Y (2015) Eye-tracking metrics in software engineering. In: 2015 Asia-Pacific software engineering conference (APSEC). IEEE Computer Society, Los Alamitos, CA, USA, pp 96–103. https://doi.org/10.1109/APSEC.2015.53. https://doi.ieeecomputersociety.org/10.1109/APSEC.2015.53
53. Sharafi Z, Sharif B, Guéhéneuc YG, Begel A, Bednarik R, Crosby M (2020) A practical guide on conducting eye tracking studies in software engineering. Empir Softw Eng 53:25
54. Silver B, Richard B (2009) BPMN method and style, vol 2. Cody-Cassidy Press, Aptos
55. Stiehl V, Danei M, Elliott J, Heiler M, Kerwien T (2019) Effectively and efficiently implementing complex business processes–a case study. In: Lübke D, Pautasso C (eds) Empirical studies on the development of executable business processes. Springer, Berlin, pp 33–57
56. Störrle H (2011) On the impact of layout quality to understanding UML diagrams. Proceedings of IEEE Symposium on Visual Lang Hum Centric Comput 2011:135–142
57. Störrle H (2014) Model-driven engineering languages and systems: 17th international conference, MODELS 2014, Valencia, Spain, September 28–October 3, 2014. Proceedings, chap. On the impact of layout quality to understanding UML diagrams: size matters. Springer, Cham, pp 518–534
58. Tufte ER (2001) The visual display of quantitative information, vol 2. Graphics Press, Cheshire, CT
59. Turetken O, Dikici A, Vanderfeesten I, Rompen T, Demirors O (2019) The influence of using collapsed sub-processes and groups on the understandability of business process models. Bus Inf Syst Eng 59:62
60. 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. Springer, pp 289–307
61. Van Solingen R, Basili V, Caldiera G, Rombach HD (2002) Goal question metric (GQM) approach. In: Marciniak JJ (ed) Encyclopedia of software engineering. Wiley, New York
62. Vanderfeesten I, Erasmus J, Traganos K, Bouklis P, Garbi A, Boultadakis G, Dijkman R, Grefen P (2019) Effectively and efficiently implementing complex business processes—a case study. In: Lübke D, Pautasso C (eds) Empirical studies on the development of executable business processes. Springer, Berlin, pp 109–134
63. Vega-Márquez OL, Chavarriaga J, Linares-Vásquez M, Sánchez M (2019) Requirements comprehension using BPMN: an empirical study. In: Lübke D, Pautasso C (eds) Empirical studies on the development of executable business processes. Springer, Berlin, pp 85–111
64. Wahn B, Ferris DP, Hairston WD, König P (2016) Pupil sizes scale with attentional load and task experience in a multiple object tracking task. PLOS ONE 11(12):1–15. https://doi.org/10.1371/journal.pone.0168087
65. Weber B, Reichert M, Mendling J, Reijers HA (2011) Refactoring large process model repositories. Comput Ind 62(5):467–486
66. White SA (2004) Process modeling notations and workflow patterns. Technical report, Object Management Group
67. Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2012) Experimentation in software engineering. Springer, Berlin
68. 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. Springer, pp 153–168

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