Received: 25 October 2022 | Revised: 3 April 2023 | Accepted: 3 July 2023

DOI: 10.1002/smr.2599

RESEARCH ARTICLE - EMPIRICAL

Journal of Software: Evolution and Process WILEY

On the perception of graph layouts

Lisa Grabinger | Florian Hauser | Jürgen Mottok

Laboratory for Safe and Secure Systems (LaS^3), Technical University of Applied Sciences Regensburg, Germany

Correspondence
Lisa Grabinger, Laboratory for Safe and Secure Systems (LaS^3), Technical University of Applied Sciences Regensburg. Germany.
Email: lisa.grabinger@oth-regensburg.de

Funding information
Bayerisches Staatsministerium für Wirtschaft, Landesentwicklung und Energie (STMWI), Grant/Award Number: DIK0173/03; Bundesministerium für Bildung und Forschung (BMBF), Grant/Award Numbers: 16DHBKI035, 13FH101IN6

Abstract

In the field of software engineering, graph-based models are used for a variety of applications. Usually, the layout of those graphs is determined at the discretion of the user. This article empirically investigates whether different layouts affect the comprehensibility or popularity of a graph and whether one can predict the perception of certain aspects in the graph using basic graphical laws from psychology (i.e., Gestalt principles). Data on three distinct layouts of one causal graph is collected from 29 subjects using eye tracking and a print questionnaire. The evaluation of the collected data suggests that the layout of a graph does matter and that the Gestalt principles are a valuable tool for assessing partial aspects of a layout.

KEYWORDS
causal graphs, eye tracking, gestalt principles, graph layouts, modeling languages

1 | INTRODUCTION

In software engineering (SE) there are various applications for graphical models, such as[1]:

- *during development or testing*: fault trees for accessing reliability in functional security[2,3] or attack trees for detecting threats to IT security.[4]
- *for specification or documentation*: class or sequence diagrams for communicating the structure or behavior of code.[5]
- *for reengineering*: program or data flow diagrams for understanding barely documented code.[6]

Most graphical models are graph-based; they consist of two general types of elements, nodes, and edges (see Figure 1A for an illustration and particular references[7,8] for a more formal introduction). Thereby, information is transferred through the selection and combination of model elements. Take again the model in Figure 1A as an example: it consists of three nodes (named X, Y, and Z) as well as two edges. The edges connect the nodes X and Z with Y, in each case toward Y. The alignment of the node elements relative to each other (i.e., the layout of the graph) does not yield any information – the graph in Figure 1B encodes the exact same information as the one in Figure 1A. In practice, the layout of the graph is determined at the user's discretion, possibly by using some sort of graph sorting algorithm.

Our research starts at this very point – combating arbitrariness in the selection of graph layouts: with an empirical study, we investigate *how*, if at all, different alignments of model elements influence the comprehensibility and popularity of a graph and *whether* we can predict the

Statements— This article is an extension of work originally presented in the "1st Workshop on Advances in Human-Centric Experiments in Software Engineering" (HUMAN 2022) [1]. We adhere to the Wiley standards for ethics and integrity. For our study, we followed the procedure recommended by the 'Joint Ethics Committee of the Universities of Bavaria' (GEHBa) responsible for our university and performed a self-assessment; we also had our subjects sign an informed consent form. The data collected in our study and the corresponding study materials are available at www.doi.org/10.5281/zenodo.7241097. Our work is funded by the 'Bavarian State Ministry of Economic Affairs, Regional Development and Energy' (STMWI) within the funding project HolmeS^3 (FKZ: DIK0173/03) and by the 'German Federal Ministry of Education and Research' (BMBF) within the funding projects HASKI (FKZ: 16DHBKI035) and FH-Invest (FKZ: 13FH101IN6). Further, our work does not reproduce material from other sources or provoke a conflict of interest, nor is it a clinical trial requiring further registration.

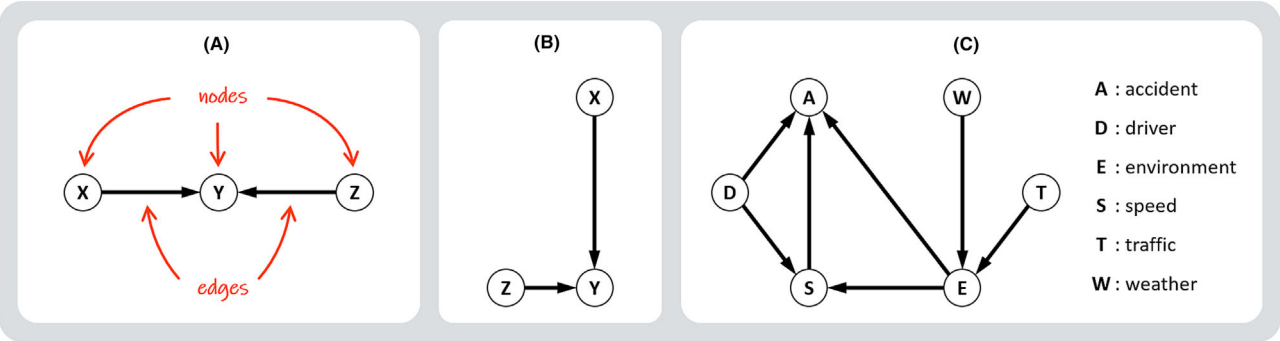

FIGURE 1 Terminology (A) and alignment variety (B) of graph-based models and an exemplary causal graph (C).

perception of certain aspects in the graph with basic graphical laws from psychology (i.e., Gestalt principles). As a research method, we rely on the triangulation of eye tracking and questioning – this allows us to access both, subconscious cognitive processes and conscious attitudes of our subjects. Causal graphs are chosen as objects of investigation because they seem to be the most intuitive.

The following pages present the conducted research starting with some background information: an explanation of the underlying theoretical constructs (i.e., causal graphs and Gestalt principles), an introduction of the research method of eye tracking, and an overview of related work. Afterwards, the study itself is described – its design, implementation, and analysis. To bring this article to a close, the results of our research are presented and discussed. Note that this article is an extension of work originally presented in a conference paper.[1]

2 | BACKGROUND

This section provides some background information on the study content to facilitate its understanding. First, the causal graphs are introduced. This is followed by the explanation of selected Gestalt principles. Then, the research method of eye tracking is presented. In the end, related approaches are put in place.

2.1 | Causal graphs

Causal graphs are graph-based models used to visually represent cause-and-effect relations between stochastic random variables by directed edges[9]– if the graphs in Figure 1A and B were interpreted as causal, they would state that the random variables X and Z *cause* the random variable Y directly (i.e., X and Z are *direct causes* of Y).[8] Causality prohibits undirected relations, bidirectional relations, or cycles – an effect cannot influence itself or its cause.[10]

For the empirical investigation, a concrete graph is needed. Based on previous work by the authors, a highly simplified model of how a car accident occurs is chosen – outlined in Figure 1C. The model states that an accident (random variable A) may happen because the driver (random variable D) is distracted or tired, the current speed of the car (random variable S) makes it impossible to brake in time, or the environmental conditions (random variable E) such as visibility or road condition are poor. It is assumed that the speed of the car is mainly influenced by its driver and the environmental conditions – the speed tends to increase once the driver is stressed and decreases when stuck in a traffic jam. In other words, environmental conditions include both, weather (random variable W) and traffic (random variable T).

2.2 | Gestalt principles

Gestalt principles are law-like observations about human perception of graphical objects. Their beginnings trace back to the early 20th century: with empirical research, the psychologist Wertheimer identified multiple factors that determine the perception of grouping.[11] For the purposes of this article, we will focus on three principles known as proximity, similarity, and closure. They state that elements are perceived to belong together if they are:

- *proximity*: … close to each other,
- *similarity*: … similar to each other, or
- *closure*: … in combination suggest a closed form.[12]

See Figure 2A-2C for common examples: a collection of dots is perceived column-wise when properly arranged (Figure 2A) or colored (Figure 2B); an implied circle is perceived as full (Figure 2C).[12] The Gestalt principles can also be applied to graphs; Figure 2D-2F shows the assumed effect of the three Gestalt principles on the causal graph from Figure 1C. Proximity suggests that closely located nodes are perceived as a unit – regardless of them being connected by edges (Figure 2D). Similarity strikes once several edges are similar to each other in that they point in a similar direction (Figure 2E). The principle of closure should show if the arrangement of model elements resembles a known form (Figure 2F).

2.3 | Eye tracking

Eye tracking is the recording of eye movement data; in current systems, this works non-invasively with nothing more than a small camera (and sometimes a light source) pointed at the subject.[13] The raw eye tracking data is simply a time series of two-dimensional coordinates[13] – the point in space and time the subject looked at. This time-series-data is usually transformed into more complex metrics mostly based on:

- *fixations*: … moments when the subject's gaze is roughly focused on some region while processing information within that region,[14]
- *saccades*: … moments when no information is taken in by the subject because the subject's gaze alternates between two areas,[14] or
- *areas of interest* (AOI): … sub-areas of the presented stimulus.[13]

The acquired data is valuable for a variety of use cases: from support systems (e.g., in the form of driver assistance systems[15] or interactive learning systems[16]) and medical applications (e.g., for gaze interaction[17] or visual acuity measurement[18]) to applied research (e.g., of usability[19] or behavioral patterns[20]).

2.4 | Related work

There is already some eye tracking-based research on our first starting point of research, the comprehensibility of graph layouts. For example, Körner[21] used eye tracking to set up a heuristic model of graph comprehension for so-called hierarchical graphs (i.e., graph-based models with undirected edges where the relative vertical position of nodes indicates the direction of their relationship); the subjects' task was to determine whether a certain directed relationship exists between two nodes a graph. The results suggest that the subjects' behavior can be divided into three stages: searching for the first node, searching for the second node, and reasoning about their relation.

Huang[22] validated the proposed three-stage model for undirected graphs by investigating the effects of individual aspects of graph layout (i.e., edge crossings and geodesic path tendency) on graph comprehensibility. The subjects' task was to look for a given number of edges – either between two nodes or toward a node. For edge crossings, Huang found a difference in performance for general edge crossings, but in the eye tracking data only for small crossing angles. However, the eye tracking data suggests a tendency toward geodesic paths.

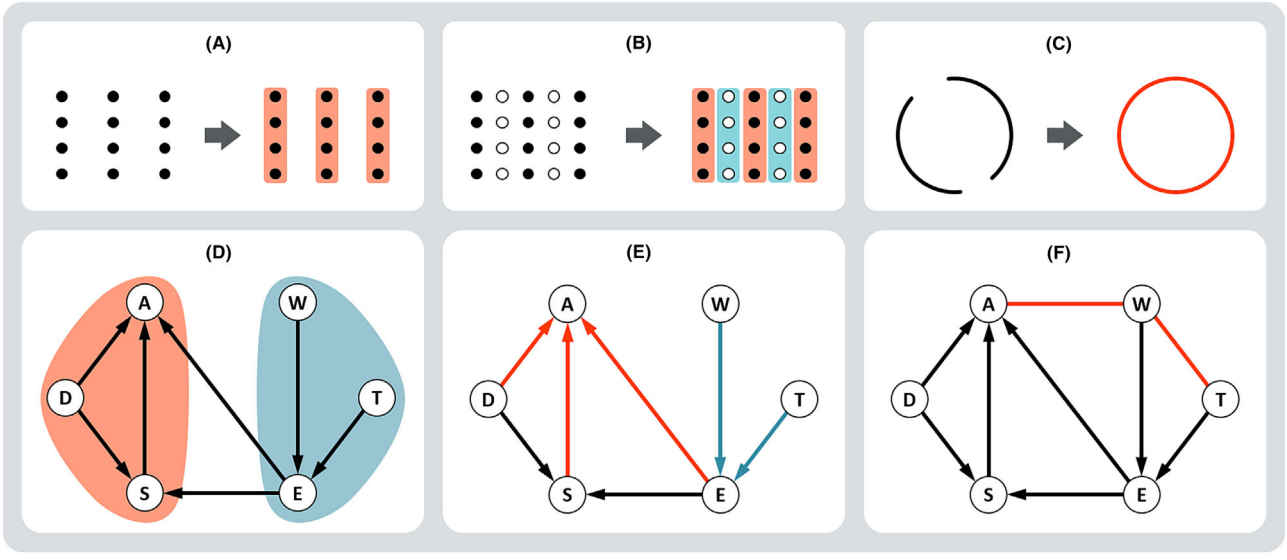

FIGURE 2 Exemplary effects of gestalt principles proximity (A, D), similarity (B, E), and closure (C, F). Based on Gerrig[12] (A-C).

There are also eye tracking studies that look at the comparison of specific layouts: for example, Pohl, Schmitt, and Diehl[23] compared three graph layouts (i.e., orthogonal, force-directed, and hierarchical) using some random undirected graphs. Again, the subjects encounter tasks based on graph theory. The results of the study indicate that the forced-directed layout is the best in terms of response time and accuracy.

In addition, some empirical studies have been conducted on diagrams entirely without eye tracking: Sharif and Maletic[24] studied different layouts (i.e., multi-cluster and orthogonal) of class diagrams (i.e., partially directed graphs) using an online questionnaire. Here, subjects completed tasks more closely related to the SE domain, such as bug fixing. The results are in favor of the multi-cluster layout.

The list of empirical work on graphs could be continued, but to our knowledge, all of them differ from the work presented in this paper in at least one of the following aspects, as the approaches presented[21–24] exemplarily show:

- *goal*: As in the latter two approaches,[23,24] the goal of this paper is to compare specific layouts. In contrast the first approach[21] works on building a heuristic model for graph understanding, while the second one[22] evaluates the impact of certain layout aspects.
- *object of investigation*: The first approach[21] studies hierarchical graphs (i.e., graphs that encode directed relationships between node elements by their layout), while the middle two[22,23] use undirected graphs (i.e., graphs that encode only undirected relationships between node elements). The present study considers directed graphs (i.e., graphs that encode directed relationships between node elements by directed edges), a concept that is far more natural for the SE domain than hierarchical or undirected graphs.
- *tasks*: In the first three approaches,[21–23] the subjects' tasks are built upon graph theory (e.g., dealing with paths, cliques, or degrees). The present study uses tasks that are more natural to the discipline of SE (i.e., memorizing and debugging or reproducing, respectively) – in line with the work done in the last apporach.[24]
- *analyses*: In the first three approaches[21–23] as well as in the present study, the different graphs are evaluated with eye tracking (e.g., heat maps or fixation counts) or without (e.g., answer times, error rates, or questioning) – in the last one[24] no eye tracking is employed. However, all of them only use quantitative data to rate the entire graph and qualitative data, if at all, to focus on distinct parts of the graph (e.g., with heatmaps). In contrast, we relate quantitative data to individual model elements. Moreover, we extend the work of all previous approaches presented[21–24] by not only collecting and evaluating empirical data but also using psychology (i.e., Gestalt principles) to make predictions in advance.

The second starting point of our research, the applicability of those Gestalt principles was investigated several times over the last 100 years. Hu and Bačić[25] even used eye tracking as a research method. However, with or without eye tracking, up to now, simple geometric arrangements (e.g., Figure 2A-2C) were chosen as object of investigation instead of more complex graphical objects such as graphs (e.g., Figure 2D-2F).

3 | METHODS

This section takes a closer look at the study conducted; thereby, the structure of the section is specifically adapted to the study that was conducted. Before the concrete hypotheses are formulated, the materials, procedure, and assumptions are discussed. Then, the collection and analysis of the study data is presented.

3.1 | Material

Since our main goal is to investigate the influence of the layout of a graphic on its perception, the independent variable (IV) is layout. Here, we consider three different values:

- *IV1 (Top-Down)*: This layout corresponds to the way a graph is usually presented in causal literature[9] – all edges point either downwards or horizontally.
- *IV2 (Bottom-Up)*: This layout is oriented opposite to IV1, with all edges pointing upwards or horizontally. This type of representation is most common in technical domains (e.g., SE), for example in so-called fault trees.[2,3]
- *IV3 (random)*: The aim of the final layout is to create a counterpart to the two tree-like structures. Here, the graph is not aligned in one direction but rather resembles a hexagonal or circular object.

Figure 3A-3C shows the individual layouts for the exemplary graph of the study presented earlier in Figure 1C. The remaining panels in Figure 3 show other graphs that the subject encounters in the course of the study – the alienated (3d-3 f) and manipulated versions (3 g-3i) of the respective layouts or *original graphs*. These are required by some tasks of the study: each subject has to memorize and reproduce one graph from all three layouts.

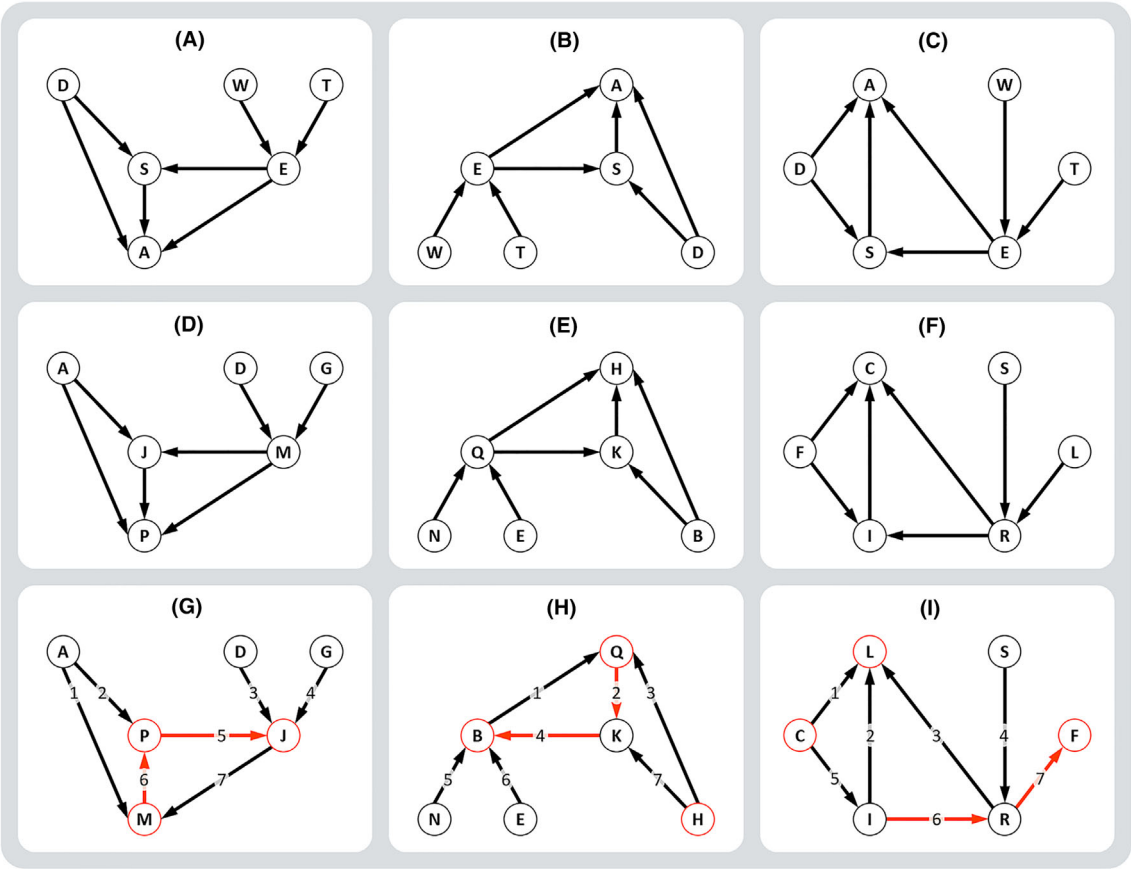

FIGURE 3 Original (A-C), alienated (D-F), and manipulated graphs (G-I) for IV1 (A, D, G), IV2 (B, E, H), and IV3 (C, F, I).

If the graphs from Figure 3A-3C were chosen for memorization and the subject is supposed to reproduce the contents of one graph, the subject could fall back on his knowledge of the other layouts for this purpose. Therefore, the graphs are alienated: the actual random variables are cyclically replaced by letters of the alphabet. To create as equal conditions as possible, exactly one vowel is assigned to each IV (which results in the vowel "o" not being assigned at all). The distribution of the assigned letters in the graphs does not follow a fixed scheme. Rather, the arrangements are chosen to minimize memory aids – for example, the spatial proximity of letters that produce common abbreviations is avoided. The resulting *alienated graphs* are shown in Figure 3D-3F.

The reproduction is realized partly as "debugging" in a manipulated version of the alienated graph. To ensure comparability between layouts, the respective manipulations must be of the same kind – we opted for the inversion of two edges and the rotation of three nodes for each layout. The resulting *manipulated graphs* are compiled in Figure 3G-3I with the altered model elements highlighted in red and the edges numbered consecutively. The former is only done for better readability of this very figure and not in the study itself, the latter serves for better addressing of particular edges in the present article as well as in the empirical study.

3.2 | Procedure

The study follows a within-subject design; each subject is exposed to each task or variation of the IV.[14] In the beginning, the subjects complete some paperwork. They sign a *consent form* and fill the first part of the *questionnaire* with their demographic data (i.e., age, gender, occupation, and area of expertise) as well as their previous experience with fault trees or causal graphs, respectively.

To increase the accuracy of eye tracking data, a *9-point-calibration* is performed for each subject prior to the actual recording: 9 dots are presented on the eye tracker's monitor one after the other with the subjects instructed to focus the particular dots with their gaze.[14]

The first stimulus of the eye tracking study includes a written *introduction* to causal graphs – similar in content to the first paragraph in section 2.1. While viewing this stimulus, the subjects have the opportunity to ask questions about the topic. This ensures that all subjects have the necessary prior knowledge and that we do not have to restrict participation in any way. After that, the actual eye tracking *data collection* follows: divided into two experiments regarding comprehensibility and popularity, respectively.

In the *first experiment*, the subjects are presented with one of the alienated graphs and instructed to memorize it for a self-selected period of time; they are then asked to name the direct causes of one node (i.e., the *reproducing task*) and to identify the changes in the corresponding manipulated graph (i.e., the *debugging task*) – all from memory and verbally, with answers noted in the questionnaire by the study leader. The experiment is repeated for all three layouts or IVs, with a trial run at the beginning. The trial run is based on an arbitrary graph constructed from six edges and five nodes with names for the random variables that do not appear in the aliased graphs (i.e., the *training graph*). The order of the three real runs is varied between the subjects as first (i.e., IV1 → IV2 → IV3), second (i.e., IV2 → IV3 → IV1), and third timeline (i.e., IV3 → IV1 → IV2). The different timelines counteract possible learning effects of the subjects: even if each subject achieves the best personal results in the last run of the trial, this effect balances out across all subjects. This ensures an objective evaluation of the comprehensibility of the individual graphs.

In the *second experiment*, some preferential looking tasks (PLTs) are performed: the subjects are presented with stimuli showing two elements of one kind (i.e., two causal graphs) side by side and instructed to view them at will.[26] Each PLT is preceded by a stimulus showing a small cross for a few seconds – the subjects are asked to focus on this cross. This ensures that the subjects' gaze on the PLT stimulus starts at a predefined position. In total, each subject encounters 18 PLTs or three runs of the experiment, respectively: the trial run consisting of six PLTs between one of the alienated graphs and the training graph, the first run with two of the alienated graphs each, and the second run with two of the original graphs each. The individual stimuli of the first and second run each follow the scheme: IV1 against IV2, IV2 against IV3, IV3 against IV1, IV2 against IV1, IV3 against IV2, and IV1 against IV3. Before the second run, the *causal story* behind the original graphs is explained to the subjects in written form similar in content to the second paragraph in section 2.1 of the present article.

The study ends with a short *retrospective interview*. There, subjects are asked about their memorizing strategy and their preference between the three layouts – before knowing the causal context (i.e., with the alienated graphs) and after learning it (i.e., with the original graphs). This choice is again noted in the questionnaire by the study leader. Also, during the interview, the subjects have the opportunity to view their gaze recording and make further comments on the study.

To give a better idea of the individual eye tracking stimuli, the sequence of the two experiments is shown in Figure 4. There, it is also noted whether the change between stimuli is triggered by a timer or by an action (i.e., mouse click) of the subject.

3.3 | Assumptions

The study design carries two main assumptions. First, with the Gestalt principles we can predict the subjects' behavior while memorizing the graph and facing the reproducing or debugging tasks. Second, with PLTs we can expose the subjects' subconscious decisions between two elements of one kind. This section explains how we reached our testable implications based on these two main assumptions and some other minor ones.

In *memorization*, we generally assume that a subject's gaze is directed along the edges of the graphs, regardless of their direction. However, Gestalt principles suggest certain deviations from the edges – in particular, that the subject's gaze:

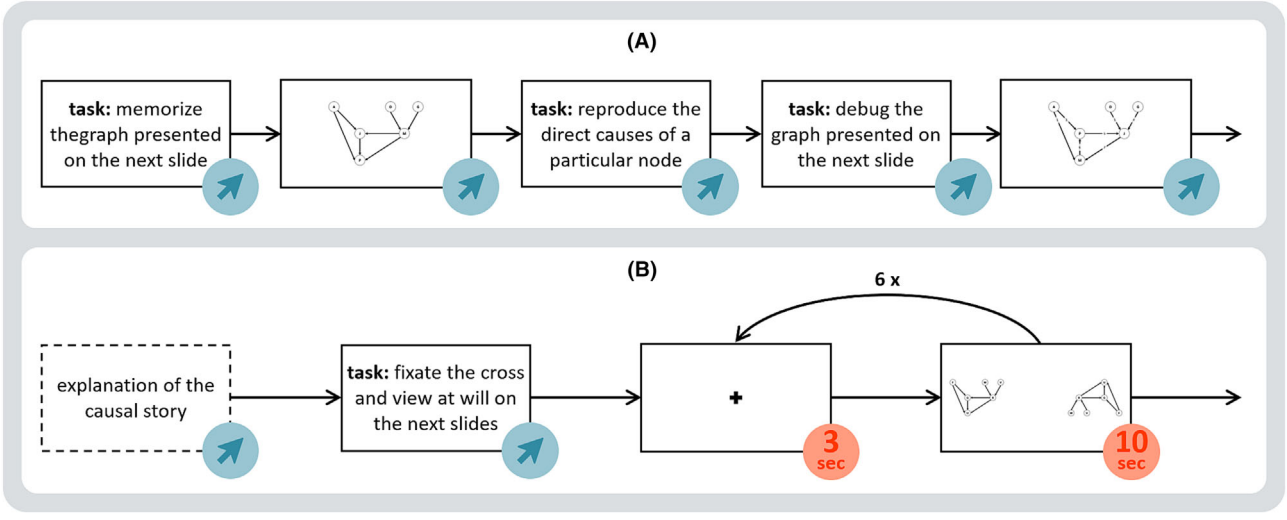

FIGURE 4 Procedures of the first (A) and second (B) experiment.

- *proximity*: … switches between nodes that are close to each other, even if there is no direct connecting edge between them (e.g., the subject's gaze should wander between the nodes D and G in the alienated graph of IV1),
- *similarity*: … preferably follows the shorter one in case of nearly parallel edges (e.g., the edge between nodes B and K should be preferred over the edge between the nodes B and H in the alienated graph of IV2), and
- *closure*: … follows the boundary lines of a shape indicated by the positioning of the nodes, even if this shape is not completely described by edges (e.g., the subject's gaze should wander between the nodes C and S in the alienated graph of IV3).

Table 1 summarizes those implications for the three layouts. Thereby, it introduces the term *transition*. For the purposes of this article, a transition will refer to the direct change of a subject's gaze between two nodes regardless of direction. When a subject's gaze wanders from node D to node G without passing through any other node, this is interpreted as a transition between D and G, but so is a change of gaze from node G to node D without passing through any other node.

The *reproducing tasks* are chosen in a way that two random variables are correct each – one from each of the two mental groupings according to proximity. We assume that it is easier for the subject to name the cause that is assigned to the same grouping as the node of the question. Consequently, the relative frequency of this answer should be greater than that of the other answer across all subjects. Figure 5A-5C outlines these considerations using IV3 as an example – once again, all edges are numbered for an easier referencing. In Figure 5A, the mental groupings of the graph relevant to the task are colored. In Figure 5B the correct answers (i.e., the direct causes of the node I) are highlighted in red. Figure 5C combines the two previous illustrations. It can be seen that the node of the quest ion (I) and one of the causes (F) belong to the red grouping, while the second cause (R) belongs to the blue grouping. The connection of nodes F and I (edge 5) falls into one grouping; the connection of nodes R

TABLE 1 Predictions for memorization.

Layout	Transition	Expected deviation	Underlying gestalt principle(s)
IV1	D ↔ G	Amendment	Proximity
IV1	A ↔ P	Omission	Similarity
IV2	E ↔ N	Amendment	Proximity
IV2	B ↔ H	Omission	Similarity
IV3	L ↔ S	Amendment	Proximity, Closure
IV3	C ↔ S	Amendment	Closure

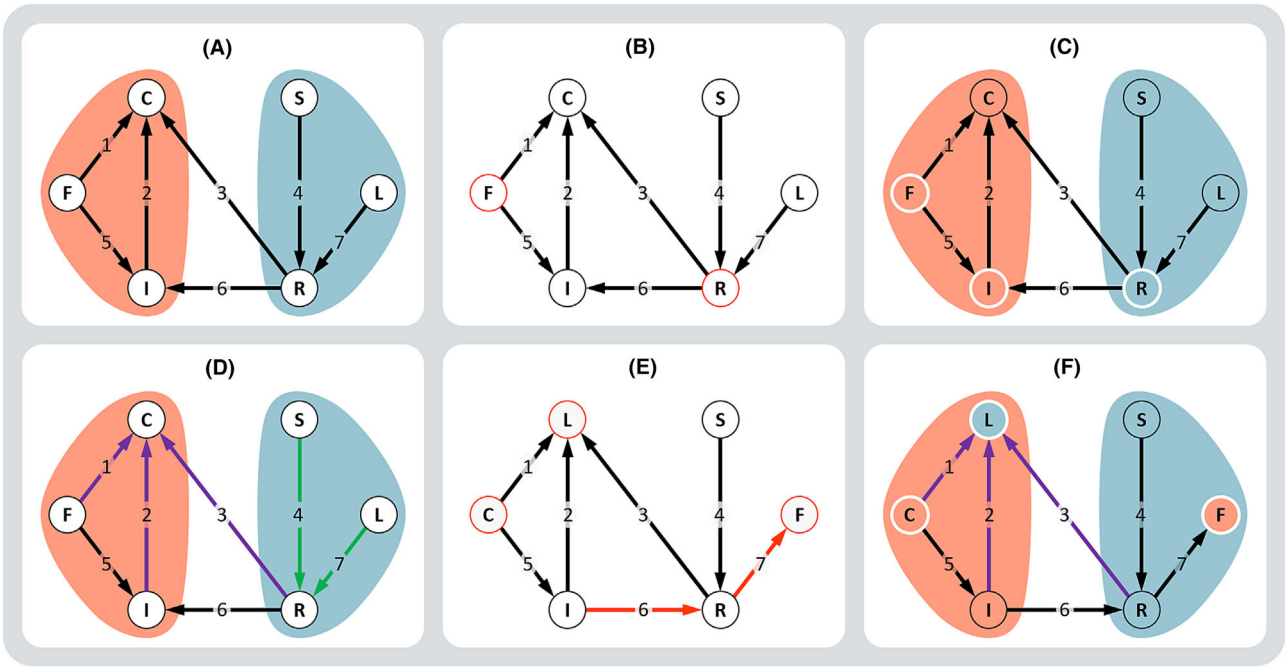

FIGURE 5 Grouping according to the gestalt principles (A, D), solution (B, E), and explanation (C, E) for reproducing (A-C) and debugging tasks (D-F) of IV3.

and I (edge 6) is in between groupings. The connection to node F should mentally be more closely grouped with the node I and thus more easily reproducible the connection to node R. The relative answer frequencies should thus satisfy the inequality P("F") ≥ P("R").

Likewise, in the *debugging tasks*, some manipulations should be easier to detect than others: we assume that switching random variables between mental groupings according to proximity should be more noticeable to subjects than switching within. Also, reversed edges should attract more attention if they break a grouping based on the similarity. Thus, again, the associated answers should occur more frequently across all subjects. Figure 5D-5F outlines this intuition for IV3. Figure 5D presents the mental groupings of the memorized graph. In Figure 5E the correct answers (i.e., the manipulations) are highlighted in color. Figure 5F once again combines the previous illustrations. The color highlighting makes the manipulation of nodes L and F stand out more than the one of C – recalling the memorized graph should produce the same effect. Same applies to the edges: by reversing edge 7, the latter grouping is missing from the manipulated graph. This gives the subjects a "clue" for detecting the manipulation of edge 7. These considerations lead to the following inequalities: P("F") ≥ P("C"), P("L") ≥ P("C"), and P("7") ≥ P("6").

Table 2 summarizes the previous explanations for the three layouts. The contents for IV1 and IV2 come without further derivation – the underlying considerations follow the same scheme as for IV3.

In the course of *PLTs* we rely on two aspects: which of the elements was looked at first and which was looked at longer by the subject. In accordance with Behe, Campbell, and Khachatryan,[26] we assume that the element that was viewed first caught the subject's attention, while that one that was viewed more extensively fascinates the subject more. This means that a PLT tells which of the given items is (subconsciously) chosen by the subject – while being more robust than questioning.[26]

In comparing the three layouts or IVs, we believe that the tree-like arrangements (i.e., IV1 or IV2) will prove advantageous in terms of comprehensibility and popularity – simply because they appear more orderly. We also assume that IV1 is more appropriate than IV2, as this is consistent with the natural way of causal thinking that leads from causes to effects, rather than the other way around. Regardless of whether PLTs or surveys are used, we firmly believe that the classification of popularity will become clearer once the causal context is known.

3.4 | Hypotheses

In accordance with section 1, we define our research questions to be 1) Do the Gestalt principles hold true for (causal) graphs? 2) Do different alignments influence the comprehensibility of a (causal) graph? 3) Do different alignments influence the popularity of a (causal) graph? For a concise evaluation, we convert those research questions into a set of nine hypotheses – with three hypotheses each grouped by an overarching formulation:

H1. The Gestalt principles of proximity, similarity, and closure hold true for causal graphs.
H1a. The transitions during memorization comply with edges or the predictions in Table 1.
H1b. The relative frequencies of correct answers to the reproducing tasks satisfy the predictions in Table 2.
H1c. The relative frequencies of correct answers to the debugging tasks satisfy the predictions in Table 2.

TABLE 2 Predictions for reproducing and debugging tasks.

Layout	Task	Solution	Expected ratio of relative answer frequencies	Underlying gestalt principle
IV1	Reproducing task for J	A, M	P("A") ≥ P("M")	Proximity
IV1	Debugging task	J, M, P, 5, 6	P("J") ≥ P("P")	Proximity
			P("M") ≥ P("P")	Proximity
			P("6") ≥ P("5")	Similarity
IV2	Reproducing task for K	B, Q	P("B") ≥ P("Q")	Proximity
IV2	Debugging task	B, H, Q, 2, 4	P("B") ≥ P("H")	Proximity
			P("Q") ≥ P("H")	Proximity
			P("2") ≥ P("4")	Similarity
IV3	Reproducing task for I	F, R	P("F") ≥ P("R")	Proximity
IV3	Debugging task	C, F, L, 6, 7	P("L") ≥ P("C")	Proximity
			P("F") ≥ P("C")	Proximity
			P("7") ≥ P("6")	Similarity

H2. The comprehensibility of causal graphs decreases from IV1 to IV2 to IV3.

H2a. The duration of memorization increases from IV1 to IV2 to IV3.

H2b. The score on the reproducing tasks decreases from IV1 to IV2 to IV3.

H2c. The score on the debugging tasks decreases from IV1 to IV2 to IV3.

H3. The popularity of the causal graphs decreases from IV1 to IV2 to IV3.

H3a. With PLTs, one chooses IV1 over IV2 over IV3.

H3b. With direct questioning, one chooses IV1 over IV2 over IV3

H3c. With the knowledge of the causal context, the effects from H3a and H3b strengthen.

Note that the quantity *score* mentioned in the hypotheses H2b and H2c is measured as a percentage, but is not the same as the percentage of correct answers – rather, the correct and incorrect answers are offset. A score of 100% can only be achieved with exactly the correct answers (each two for the reproducing and five for the debugging tasks as listed in Table 2); each wrong or missing answer leads to point deduction of 50% for the reproducing tasks and 20% for the debugging tasks.

Figure 6 breaks down the data sources used to evaluate the hypotheses – for the memorization (6a), the reproducing or debugging tasks (6b), and for the PLTs. Figure 6A and C also show the empirically chosen AOIs – as circles over the nodes of the graphs to be memorized or as rectangles over the elements of the PLTs. The transitions between the nodes during graph memorization (6a) cannot be exported directly from the eye tracker's analysis tool, but must be reconstructed from the exportable activations of the set AOIs – this procedure is further elaborated in section 3.6. In the present case, the memorization duration (6a) can be equated with the viewing duration of the corresponding stimulus – a quantity that is not actually an eye tracking metric in the strict sense, but is recorded by the eye tracker. The answers to the reproducing or debugging tasks as well as the deliberate decisions of the subjects (6b) are not assigned a metric – they are taken from the questionnaire instead of from the eye tracker. From the PLT stimuli (6c) we export two eye tracking metrics: the time span until the first gaze into each AOI and the total time duration of their viewing.

The utilized eye tracking metrics are proven to be valid and reliable.[14] Their naming complies with the utilized analysis tool of the eye tracker.

3.5 | Realization

We utilized the monitor-based eye tracker Tobii Pro Spectrum (monitor: 23.8 in.; 16:9), its associated analysis tool Tobii Pro Lab version 1.145.28180 (x64), and a print questionnaire. The eye tracking data was collected contact-free at a frequency of 300 Hz; the print questionnaire was filled in partly by the subject and partly by the study leader. The entire process took about 20 to 40 minutes per subject.

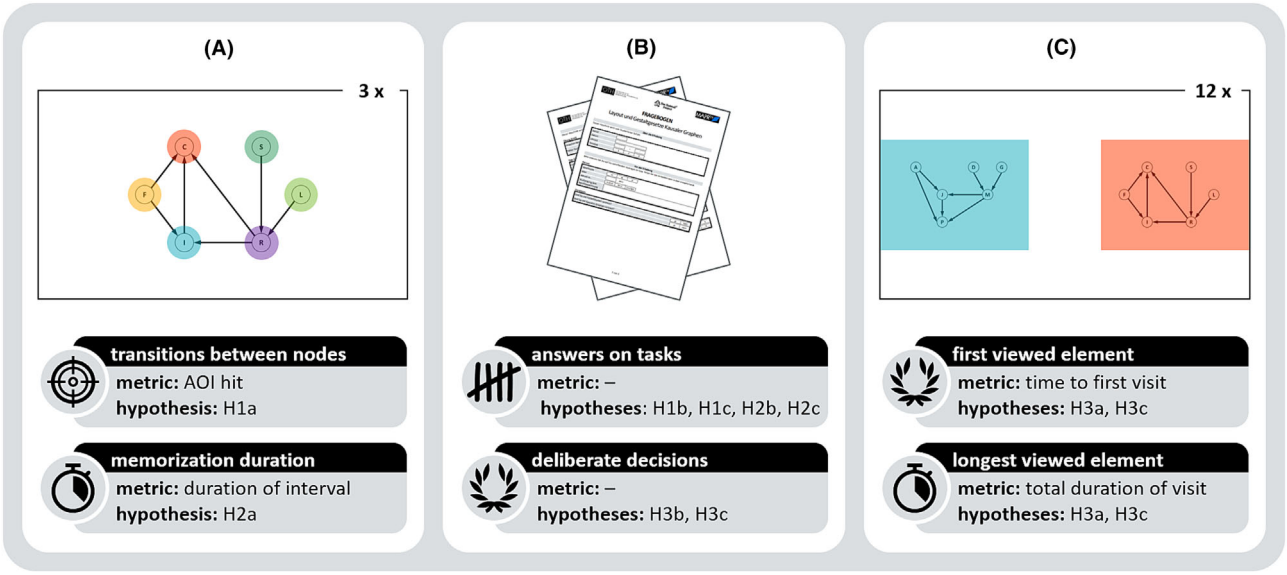

FIGURE 6 Data for the evaluation of hypotheses during memorization (A), reproducing or debugging tasks (B), and PLTs (C).

We promoted the study at the Technical University of Applied Sciences Regensburg and did not restrict participation by prior knowledge, area of expertise, or any other factor. Thus, we recruited 29 subjects aged 21 to 64 years (median = 25). Among them, 11 were women (38%). At the time of the study, most subjects were enrolled in a university (24 subjects, 83%) and worked in STEM fields (22 subjects, 76%).

3.6 | Analysis

The data collected was analyzed with the programming language R version 4.2.1 using the libraries *base*[27] and *rstatix*.[28] Data analysis is quantitative throughout – however, the first hypothesis (i.e., H1) is examined using descriptive statistics only, while the other hypotheses (i.e., H2 and H3) are examined with inductive statistics in terms of statistical tests.

For hypothesis H1a we start by exporting the metric AOI hit from the analysis software as one data set per subject, each of which contains one column per AOI or node. These columns are filled with one value per measuring point:

- "NA" if the stimulus was not presented at the corresponding measuring point,
- "1" if the AOI was activated (i.e., viewed) at the corresponding measuring point, and
- "0" if the AOI was not activated (i.e., not viewed) at the corresponding measuring point.

To get from those data sets to our quantity of interest (i.e., the number of transitions between every two nodes during memorization) we repeat a 5-step-process for every subject and layout – visualized in Figure 7 for a highly simplified exemplary gaze on IV3 (i.e., C → F → I → C → S → L → R → S → R → I → F). As a first step, we restrict the dataset to the data of one layout – by deleting the columns that do not belong to the AOIs of the current layout and then removing the rows that exclusively contain the value NA (i.e., the rows that belong to measuring points when another layout was presented). In this *cleaned data set*, we determine the *rising edges* in the binary coding of each column (i.e., the entries of gaze into the corresponding AOI). This is done by subtracting each value from its predecessor; a difference of 1 marks a rising edge. In step 2, we fill an *auxiliary vector* with the column names of the cleaned dataset (i.e., the names of

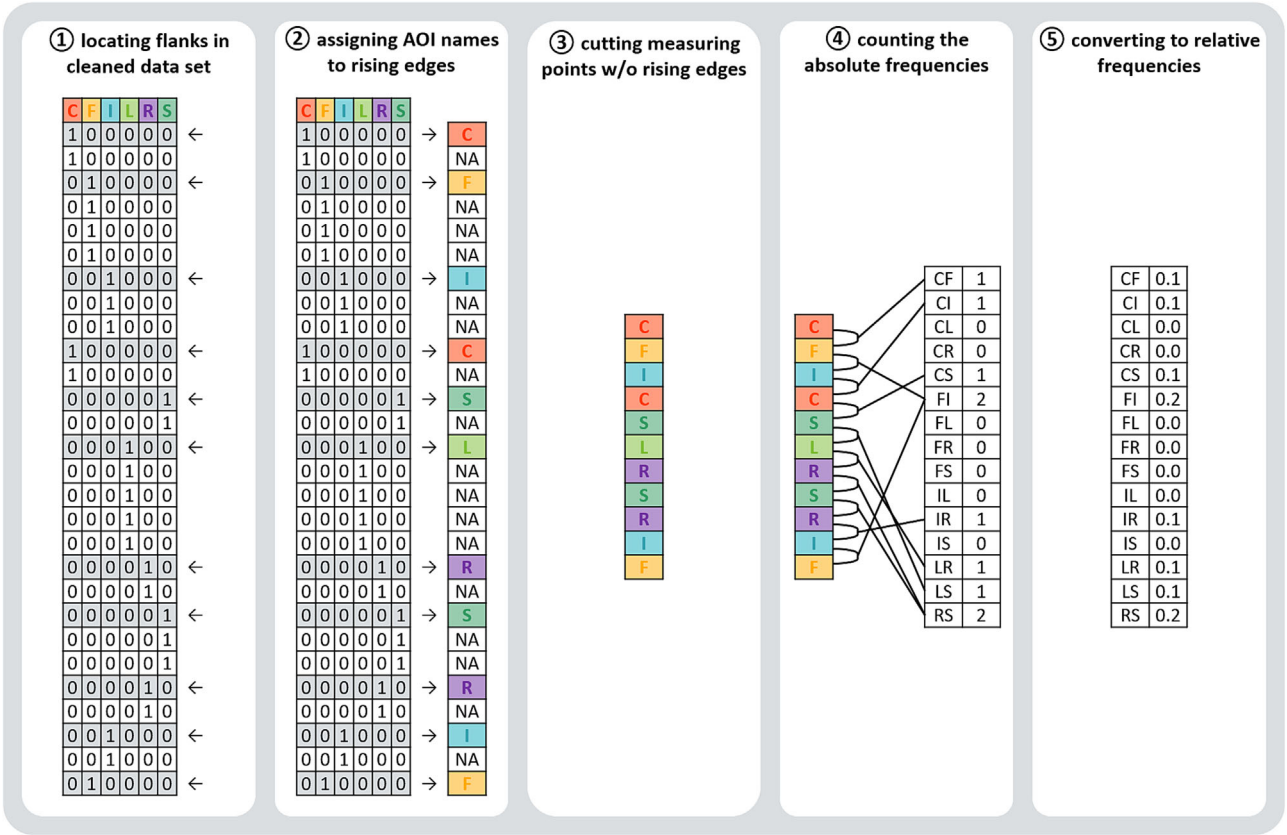

FIGURE 7 Process for determining the relative frequencies of transitions for a subject.

the AOIs) at the indices that were found to be rising edges in the respective column. Removing the rows without rising edges from the auxiliary vector (i.e., step 3) gives an *ordered listing* of visited nodes. The *absolute frequencies* of transitions can be taken directly from this listing (i.e., step 4). For example, whenever C is followed by F, the absolute frequency of the transition C ↔ F is increased by 1. In a final step, we convert those absolute numbers to *relative* ones; this is necessary because the number of transitions varies greatly between subjects, ranging from a total of 22 transitions on the entire graph (i.e., subject P18) to 44 transitions only between two particular nodes (i.e., subject P29).

The evaluation of hypotheses H1b and H1c is more straightforward: we simply determine the relative frequencies with which each model element is named in the course of the reproducing or debugging tasks.

To access the hypotheses H2a, H2b, H2c, H3a, and H3b, we need to find out whether the means or medians of the respective variables differ significantly for the distinct layouts (i.e., three groups). Due to the within-subjects design of our study, the data collected is dependent between the three groups. Following Field and Hole[29] we use either a *one-way repeated measures ANOVA* for a joint comparison, followed by *dependent t-tests* for pairwise comparisons, or *Friedman's ANOVA* for a joint comparison, followed by *Wilcoxon signed-rank tests* for pairwise comparisons. The choice between these two alternatives is based on the nature of our data: if the respective variable is not normally distributed for at least one group or if sphericity between groups is not given, we choose the latter non-parametric alternative. To check for normal distribution within groups and for sphericity between groups, we rely on *Shapiro-Wilk tests* (significant ⇒ no normal distribution) and *Mauchly's tests* (significant ⇒ no sphericity), respectively.

For hypothesis H3c, we need to split the data into two groups by time: before the subjects knew the causal story and after they learned it. Once again, our data is dependent as collected from the same subjects for both of the groups. As a result, we again follow Field and Hole[29] and use *dependent t-tests* or *Wilcoxon signed-rank tests* preceded by *Shapiro-Wilk tests* for deciding between the former parametric and the latter non-parametric test alternative.

In the course of the present article, the *p-values* of *one-way repeated measures ANOVAs* are adjusted using the *Greenhouse-Geisser correction* or *Hyunh-Feldt correction* method when sphericity is not given; the choice between those two corrections is based on the computed ε (i.e., *Greenhouse-Geisser* for $\varepsilon < 0.75$, *Hyunh-Feldt* for $\varepsilon < 0.75$).[29] Meanwhile, *Bonferroni correction* is applied to all pairwise tests.[29]

All (adjusted) p-values are evaluated against a significance level of $\alpha = 0.05$. In line with Field and Hole,[29] for each test, we report the *p-value*, the *test statistic* and, when possible, the *effect size*; we do not report effect size measures for *Shapiro-Wilk tests*, *Mauchly's tests*, or pairwise comparisons.[29] For the choice of a proper effect size measure, we align with Albers and Lakens[30] for parametric and Tomczak and Tomczak[31] for non-parametric tests (i.e., generalized ε^2 for *one-way repeated measures ANOVAs*, Hedges' g for *dependent t-tests*, Kendall's W for *Friedman's ANOVA*, and the correlation coefficient r for *Wilcoxon signed-rank tests*).

4 | RESULTS

This section presents the results of the empirical study – organized according to the three research questions or main hypotheses of 1) the applicability of Gestalt principles, 2) the comprehensibility of distinct layouts, and 3) their popularity.

4.1 | Research question 1 – gestalt principles in causal graphs

Hypothesis H1a formulates the intuition that when memorizing a causal graph, the subjects' gaze tends to move along edges or deviate according to the Gestalt principles. This is accessed via the frequencies of transitions between node elements: Table 3 lists the medians of the relative frequencies of transitions across all subjects, for each layout or transition. The color highlighting of cells indicates whether the transitions are based on an edge (i.e., black cell highlighting), are made according to the Gestalt principles despite the absence of an edge (i.e., red cell highlighting), or are avoided according to the Gestalt principles despite the presence of an edge (i.e., blue cell highlighting). The individual rows in this table do not sum to 1 by definition; this is because we calculate the relative frequencies of the transitions for each subject – as explained in section 3.6 – and then median over all these percentage values. This procedure ensures to account equally for the distribution of each subject's transitions (i.e., that data from subjects such as P18 with very few transitions do not drown against data from subjects such as P29 with many transitions). To emphasize this fact, we have decided to present the median values as decimal numbers instead of percentages.

In the table, three transitions are highlighted in red: M ↔ P (IV1), D ↔ J (IV1), and B ↔ H (IV2). For these transitions, the data do not really agree with our predictions based on the Gestalt principles; the remaining transitions either have a rather small median (i.e., less than 0.02), while they are considered absent, or have a rather large median (i.e., more than 0.06), while they are considered present. This means that Table 3 partly supports hypothesis H1a.

TABLE 3 Medians of the relative frequencies of transitions.

Layout	Values														
IV1	A ↔ D	A ↔ G	A ↔ J	A ↔ M	A ↔ P	D ↔ G	D ↔ J	D ↔ M	D ↔ P	G ↔ J	G ↔ M	G ↔ P	J ↔ M	J ↔ P	M ↔ P
	0.02	0.00	0.14	0.01	0.00	0.12	0.04	0.07	0.02	0.01	0.08	0.00	0.17	0.21	0.03
IV2	B ↔ E	B ↔ H	B ↔ K	B ↔ N	B ↔ Q	E ↔ H	E ↔ K	E ↔ N	E ↔ Q	H ↔ K	H ↔ N	H ↔ Q	K ↔ N	K ↔ Q	N ↔ Q
	0.00	0.05	0.11	0.00	0.00	0.00	0.01	0.07	0.09	0.25	0.00	0.09	0.00	0.16	0.09
IV3	C ↔ F	C ↔ I	C ↔ L	C ↔ R	C ↔ S	F ↔ I	F ↔ L	F ↔ R	F ↔ S	I ↔ L	I ↔ R	I ↔ S	L ↔ R	L ↔ S	R ↔ S
	0.16	0.18	0.00	0.06	0.07	0.07	0.00	0.01	0.00	0.00	0.08	0.01	0.07	0.10	0.08

TABLE 4 Observations for reproducing and debugging tasks.

		Ratio of relative answer frequencies	
IV	Task	Expected	Observed
IV1	Reproducing task for J	P("A") ≥ P("M")	100.00% > 72.41%
IV1	Debugging task	P("J") ≥ P("P")	100.00% > 58.62%
		P("M") ≥ P("P")	100.00% > 58.62%
		P("6") ≥ P("5")	100.00% > 75.86%
IV2	Reproducing task for K	P("B") ≥ P("Q")	93.10% > 72.41%
IV2	Debugging task	P("B") ≥ P("H")	96.55% > 72.41%
		P("Q") ≥ P("H")	100.00% > 72.41%
		P("2") ≥ P("4")	100.00% > 65.52%
IV3	Reproducing task for I	P("F") ≥ P("R")	82.76% > 41.38%
IV3	Debugging task	P("L") ≥ P("C")	89.66% > 44.83%
		P("F") ≥ P("C")	100.00% > 44.83%
		P("7") ≥ P("6")	100.00% > 37.93%

Hypotheses H1b and H1c assume that answers to the reproducing or debugging tasks can be predicted through the Gestalt principles. This is evaluated with Table 4; this table enhances Table 2 with the actually observed relative answer frequencies on tasks. Here, the data strongly support all of our predictions and thus also the hypotheses.

4.2 | Research question 2 – comprehensibility of layouts

Hypotheses H2a, H2b, and H2c state that the memorizing duration as well as the scores on the reproducing or debugging tasks change between the distinct layouts: the former should increase from IV1 to IV2 to IV3, while the latter two should decrease that way. To verify this, we rely on some statistical tests, the results of which are presented in Table; there, color highlighting of cells indicates selection of the appropriate test, as explained in section 3.6. From those, we see that:

- the memorization duration is significantly lower for IV1 than for IV2 or IV3,
- the score on the reproducing tasks is significantly higher for IV1 than for IV3, and
- the score on the debugging tasks is significantly higher for IV1 or IV2 than for IV3.

We found no significant differences in the memorization duration of IV2 and IV3, in the scores on the reproducing tasks of IV2 and the other two layouts, or in the scores on the debugging tasks of IV1 and IV2. In other words: we did not find any significance contradicting the hypotheses H2a, H2b, or H2c, but we could not completely confirm them either.

TABLE 5 Statistical analyses for layouts (n = 29).

Layout	Quantity	Median	Mean	Standard deviation	Shapiro-Wilk test	Mauchly's test	Friedman's ANOVA
IV1	Memorization duration in msec	34617.00	38583.86	16818.75	W = 0.90, p = 0.01	$\chi^2(2) = 0.85$, p = 0.12	$\chi^2(2) = 19.24$, p = 0.00, W = 0.33 (medium)
IV2		39833.00	44503.48	20619.69	W = 0.91, p = 0.02		
IV3		42600.00	49418.38	23004.25	W = 0.82, p = 0.00		
IV1	Score on reproducing tasks in %	100.00	81.03	28.07	W = 0.66, p = 0.00	$\chi^2(2) = 0.83$, p = 0.08	$\chi^2(2) = 8.99$, p = 0.01, W = 0.15 (small)
IV2		100.00	77.59	34.29	W = 0.67, p = 0.00		
IV3		50.00	51.72	43.27	W = 0.78, p = 0.00		
IV1	Score on debugging tasks in %	80.00	86.21	16.13	W = 0.78, p = 0.00	$\chi^2(2) = 0.91$, p = 0.29	$\chi^2(2) = 19.39$, p = 0.00, W = 0.33 (medium)
IV2		80.00	82.76	17.50	W = 0.75, p = 0.00		
IV3		60.00	60.69	24.19	W = 0.88, p = 0.00		
IV1	Number of longest views in %	50.00	48.71	19.86	W = 0.94, p = 0.11	$\chi^2(2) = 0.96$, p = 0.61	$\chi^2(2) = 1.52$, p = 0.47, W = 0.03 (no)
IV2		50.00	49.57	22.28	W = 0.95, p = 0.20		
IV3		50.00	51.72	19.11	W = 0.93, p = 0.07		
IV1	Number of first views in %	50.00	52.16	12.97	W = 0.79, p = 0.00	$\chi^2(2) = 0.90$, p = 0.25	$\chi^2(2) = 7.97$, p = 0.02, W = 0.14 (small)
IV2		50.00	44.38	12.28	W = 0.89, p = 0.00		
IV3		50.00	53.02	9.83	W = 0.76, p = 0.00		
IV1	Number of deliberate choices in %	50.00	53.45	42.11	W = 0.79, p = 0.00	$\chi^2(2) = 0.81$, p = 0.06	$\chi^2(2) = 5.79$, p = 0.06, W = 0.10 (no)
IV2		0.00	27.59	36.81	W = 0.71, p = 0.00		
IV3		0.00	18.97	28.07	W = 0.66, p = 0.00		

TABLE 5 (Continued)

Layout	One-way repeated measures ANOVA	Wilcoxon signed-rank tests			Dependent t-tests		
		IV1-IV2	IV1-IV3	IV2-IV3	IV1-IV2	IV1-IV3	IV2-IV3
IV1 IV2 IV3	$F(2,56) = 8.33$, p = 0.00, $\varepsilon^2 = 0.04$ (small)	T = 83.50, p = 0.01	T = 54.00, p = 0.00	T = 136.00, p = 0.24	$t(28) = -2.76$, p = 0.03	$t(28) = -3.56$, p = 0.00	$t(28) = -1.81$, p = 0.24
IV1 IV2 IV3	$F(2,56) = 5.94$, p = 0.00, $\varepsilon^2 = 0.10$ (medium)	T = 59.00, p = 1.00	T = 135.00, p = 0.01	T = 160.00, p = 0.11	$t(28) = 0.44$, p = 1.00	$t(28) = 3.34$, p = 0.01	$t(28) = 2.35$, p = 0.08
IV1 IV2 IV3	$F(2,56) = 14.72$, p = 0.00, $\varepsilon^2 = 0.24$ (large)	T = 144.50, p = 1.00	T = 233.00, p = 0.00	T = 223.50, p = 0.00	$t(28) = 0.80$, p = 1.00	$t(28) = 4.87$, p = 0.00	$t(28) = 3.91$, p = 0.00
IV1 IV2 IV3	$F(2,56) = 0.11$, p = 0.90, $\varepsilon^2 = -0.03$ (no)	T = 136.00, p = 1.00	T = 150.00, p = 1.00	T = 181.00, p = 1.00	$t(28) = -0.12$, p = 1.00	$t(28) = -0.51$, p = 1.00	$t(28) = -0.32$, p = 1.00
IV1 IV2 IV3	$F(2,56) = 2.83$, p = 0.07, $\varepsilon^2 = 0.06$ (small)	T = 91.00, p = 0.24	T = 65.50, p = 1.00	T = 21.00, p = 0.07	$t(28) = 1.70$, p = 0.30	$t(28) = -0.24$, p = 1.00	$t(28) = -2.44$, p = 0.06
IV1 IV2 IV3	$F(2,56) = 4.77$, p = 0.01, $\varepsilon^2 = 0.12$ (medium)	T = 212.00, p = 0.21	T = 177.50, p = 0.02	T = 68.00, p = 0.99	$t(28) = 1.88$, p = 0.21	$t(28) = 3.02$, p = 0.02	$t(28) = 0.93$, p = 1.00

TABLE 6 Statistical analyses for context knowledge (n = 29).

Layout	Quantity	Median		Mean		Standard deviation		Shapiro-Wilk t-test		Wilcoxon signed-rank test	Dependent t-test
		w/o	w/	w/o	w/	w/o	w/	w/o	w/		
IV1	Number of longest views in %	50.00	50.00	46.55	50.86	26.49	29.49	W = 0.92, p = 0.02	W = 0.90, p = 0.01	T = 101.50, p = 0.49, r = -0.09 (no)	t(28) = 0.59, p = 0.56, g = 0.11 (no)
IV2		50.00	50.00	52.59	46.55	30.87	28.13	W = 0.91, p = 0.02	W = 0.92, p = 0.03	T = 108.50, p = 0.56, r = -0.08 (no)	t(28) = -0.84, p = 0.41, g = -0.15 (no)
IV3		50.00	50.00	50.86	52.59	26.29	25.31	W = 0.89, p = 0.01	W = 0.89, p = 0.01	T = 91.00, p = 0.82, r = -0.03 (no)	t(28) = 0.27, p = 0.79, g = 0.05 (no)
IV1	Number of first views in %	50.00	50.00	47.41	56.9	16.83	16.22	W = 0.66, p = 0.00	W = 0.79, p = 0.00	T = 67.00, p = 0.02, r = -0.30 (medium)	t(28) = 2.49, p = 0.02, g = 0.45 (small)
IV2		50.00	50.00	45.69	43.97	13.48	15.89	W = 0.61, p = 0.00	W = 0.78, p = 0.00	T = 32.50, p = 0.59, r = -0.07 (no)	t(28) = -0.57, p = 0.57, g = -0.10 (no)
IV3		50.00	50.00	56.9	49.14	16.22	10.53	W = 0.68, p = 0.00	W = 0.58, p = 0.00	T = 9.00, p = 0.06, r = -0.25 (small)	t(28) = -2.20, p = 0.04, g = -0.40 (small)
IV1	Number of deliberate choices in %	0.00	100.00	41.38	65.52	50.12	48.37	W = 0.63, p = 0.00	W = 0.60, p = 0.00	T = 40.00, p = 0.02, r = -0.30 (small)	t(28) = 2.54, p = 0.02, g = 0.46 (small)
IV2		0.00	0.00	24.14	31.03	43.55	47.08	W = 0.53, p = 0.00	W = 0.58, p = 0.00	T = 22.50, p = 0.53, r = -0.08 (no)	t(28) = 0.70, p = 0.49, g = 0.13 (no)
IV3		0.00	0.00	34.48	3.45	48.37	18.57	W = 0.60, p = 0.00	W = 0.18, p = 0.00	T = 0.00, p = 0.00, r = -0.39 (medium)	t(28) = -3.55, p = 0.00, g = -0.64 (medium)

4.3 | Research question 3 – popularity of layouts

The hypotheses H3a and H3b are similar to the previously discussed hypotheses. They state that three further quantities change between the distinct layouts: the percentage with which each layout is viewed the longest or first in the course of PLTs* and the percentage with which it is chosen deliberately in direct questioning†. All three quantities should decrease from IV1 to IV2 to IV3. Again, we resort to a series of statistical tests, the results of which are shown in Table 5. Though, we can only find one significance here: the number of deliberate choices is significantly higher for IV1 than for IV3. This observation is consistent with hypothesis H3b, yet we cannot fully confirm the hypotheses H3a and H3b.

Hypothesis H3c examines the same three quantities, but this time does not compare between layouts, but whether or not subjects knew about the causal storyline. It is hypothesized that the popularity of IV1 over IV2 over IV3 should strengthen with causal context in both, PLTs and surveys. Again, we performed statistical tests (see Table 6 for results) and found a few significances, all of which support the hypothesis:

- the number of first views of IV1 is significantly higher with causal context,
- the number of deliberate choices of IV1 is significantly higher with causal context, and
- the number of deliberate choices of IV3 is significantly lower with causal context.

5 | CONCLUSION

This section concludes the present article by discussing findings, limitations, and implications of the presented work.

5.1 | Findings

The results presented in section 4 suggest that the Gestalt principles – or at least the principles of proximity, similarity, and closure – hold for (causal) graphs. Moreover, the results showed that the layout of a causal graph affects its comprehensibility and popularity. In the actual use case, IV1 – the downwards-oriented graph – proved to be the most popular and the most suitable for memorizing, reproducing, and debugging.

In addition to this, the results of the study suggest that PLTs should be used with caution for low-emotion or recurrent elements. *First*, a PLT provides information about which of the presented elements is found more interesting by the subject – however, it remains unclear why the subject decides this way. For emotional elements, such as photos of facial expressions or vacation spots, interestingness and preference coincide. Causal graphs, on the other hand, convey little emotion; it is possible that subjects look at a causal graph longer if it seems more complex – and thus more interesting – to them. *Second*, with recurring elements, subjects had the opportunity to develop not only likes but also dislikes for certain layouts – the retrospective interviews showed that when one of the tree-like alignments was preferred (e.g., IV1), the other was strongly disliked (e.g., IV2). In a PLT, a neutrally rated element (e.g., IV3) is preferred to the disliked element (e.g., IV2) provoking bias.

5.2 | Limitations

The validity of the results presented is subject to some limitations, including:

- *the subjects*: We cannot tell whether the subjects really did their best when working on the tasks. In addition, we cannot ensure that the sample was heterogeneous enough; as described in section 3.5, most subjects were of similar age and background.
- *the independent variable*: The particular arrangement of the model elements was chosen by the authors at their discretion. We cannot assert that there is no other layout that might prove advantageous for causal graphs.
- *the material*: The stimuli present graphs with six nodes and seven edges. This is due to the study design: the subjects are asked to memorize and reproduce a given graph; a more complex graph would not allow for this type of task. However, the graphs used in everyday software engineering are much more extensive – we cannot assure that the results presented are valid for other causal graphs, let alone for other types or extensions of graph-based models.

The study materials (i.e., stimuli and questionnaire) as well as our collected data are available at www.doi.org/10.5281/zenodo.7241097 to facilitate a replication with a different sample or design.

5.3 | Implications

If one had to sum up the message of this article in one phrase, it would be: "graph layout matters" – the alignment of model elements in a graph influences their perception. This is already clear from the work of Pohl, Schmitt, and Diehl[23] or Sharif and Maletic,[24] respectively; the contribution of our work lies in the validation of this statement for directed graphs of SE, in particular for causal graphs, but also in the use of a quantitative analysis at the level of model elements. With the present article, we want to encourage

- … *practitioners* to pay attention to the design of their graphs as the perception of a graph can be improved by simply adjusting its layout. We suggest Gestalt principles as a tool to predict the understanding of certain aspects of graphs.
- … *researchers* to further investigate, which layout proves beneficial for which graph type or use case.

In future work, our findings can be used to develop style guides for graph-based modeling techniques. This way, they can be integrated into the software and system development process and support the entire SE community in their work.

ACKNOWLEDGEMENTS

We thank Dr. Rebecca Reuter, Dr. Kenneth Holmqvist, and Lena Federl for their insights and expertise that greatly supported the development and prototyping of the present study as part of an introductory eye tracking course held at the University of Regensburg.

Moreover, thanks go to the funding project FH-Invest (FKZ: 13FH101IN6) run by Prof. Dr. Jürgen Mottok for providing equipment for the eye tracking laboratory and Prof. Dr. Christian Wolff from the University of Regensburg for arranging the laboratory areas.

The present paper is supported by the 'Bavarian State Ministry of Economic Affairs, Regional Development and Energy' (STMWI) through the granting of the funding project HolmeS^3 (FKZ: DIK0173/03) and by the 'German Federal Ministry of Education and Research' (BMBF) through the granting of the funding project HASKI (FKZ: 16DHBKI035). Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Study: Layout of Causal Graphs at https://doi.org/10.5281/zenodo.7241097.

ORCID

Lisa Grabinger https://orcid.org/0000-0003-1874-6268
Florian Hauser https://orcid.org/0000-0002-2531-3295
Jürgen Mottok https://orcid.org/0000-0002-7727-2448

ENDNOTES

* Each subject encounters each layout eight times in a PLT, including four times with and without causal context.

† Each subject chooses twice, including once with and without causal context.

REFERENCES

1. Grabinger L, Hauser F, Mottok J. "Accessing the presentation of causal graphs and an application of gestalt principles with eye tracking." In *Proc. 29th IEEE Int. Conf. Software Analysis, Evolution and Reengineering (SANER 2022)*, Honolulu, HI, USA; 2022: 1267-1274. 10.1109/SANER53432.2022.00153.
2. Edler F, Soden M, Hankammer R. *Fehlerbaumanalyse in Theorie und Praxis*. Springer Vieweg; 2015.
3. *Funktionale Sicherheit sicherheitsbezogener elektrischer/elektro-nischer/programmierbarer elektronischer Systeme; Teil 7: Überblick über Verfahren und Maßnahmen*, IEC 61508-3:2011; 2011.
4. Eckert C. *IT-Sicherheit: Konzepte–Verfahren–Protokolle*. 10th ed. De Gruyter; 2018.
5. Rupp C, Queins S, the SOPHISTs. *UML 2 glasklar: Praxiswissen für die UML-Modellierung*. 4th ed. Carl Hanser Verlag; 2012.
6. Ebert J. "Software-Reengineering Umgang mit Software-Altlasten." In Proc. Informatiktage 2003, Bad Schussenried, Germany; 2003: 24-31.
7. Cormen T, Leiserson C, Rivest R, Stein C. *Algorithmen – Eine Einführung*. 2nd ed. Oldenbourg; 2004.
8. Koller D, Friedman N. *Probabilistic graphical models: principles and techniques*. MIT Press; 2009.
9. Pearl J, Mackenzie D. *The book of why: the new science of cause and effect*. Basic Books; 2018.
10. Bowers RI. Causal reasoning. In: Shackelford T, Weekes-Shackelford V, eds. *Encyclopedia of evolutionary psychological science*. Springer; 2017:1-17.
11. Wertheimer M. Untersuchungen zur Lehre von der Gestalt II. *Psychol Forsch*. 1923;4(1):301-350. doi:10.1007/BF00410640
12. Gerrig R. *Psychologie*. 21st ed. Pearson; 2018.
13. Blake C. Eye-Tracking: Grundlagen und Anwendungsfelder. In: Möhring W, Schlütz D, eds. *Handbuch Standardisierte Erhebungsverfahren in der Kommunikationswissenschaft*. Springer VS; 2013:367-387.

14. Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, Van de Weijer J. *Eye tracking: a comprehensive guide to Methods and measures*. Oxford University Press; 2011.
15. Singh S, Papanikolopoulos N. "Monitoring driver fatigue using facial analysis techniques." In *Proc*. 1999 IEEE/IEEJ/JSAI Int. Conf. *Intelligent Transportation Systems*, Tokyo, Japan; 1999: 314-318. 10.1109/ITSC.1999.821073.
16. Castner N, Kasneci E, Kübler T, et al., "Scanpath comparison in medical image reading skills of dental students: Distinguishing stages of expertise development." In *Proc*. 2018 ACM Symp. Eye Tracking Research and Applications (ETRA 2018), Warsaw, Poland; 2018: 1-9. 10.1145/3204493.3204550.
17. Hansen J, Agustin J, Skovsgaard H. "Gaze interaction from bed." In *Proc. 1st Conf. Novel Gaze-Controlled Applications (NGCA 2011)*, Karlskrona, Sweden; 2011: 1-4. 10.1145/1983302.1983313.
18. Woodhouse J, Morjaria S, Adler P. Acuity measurements in adult subjects using a preferential looking test. *Ophthalmic Physiol Opt*. 2007;27(1):54-59. doi:10.1111/j.1475-1313.2006.00454.x
19. van't Klooster J, Slijkhuis P, van Gend J, Bente B, van Gemert-Pijnen L. "First eyetracking results of dutch coronamelder contact tracing and notification app." In *Proc. 12th Int. Conf. Intelligent Human Computer Interaction (IHCI 2020)*, Daegu, South Corea; 2020: 199-207. 10.1007/978-3-030-68452-5_21.
20. Busjahn T, Bednarik R, Begel A, et al. "Eye movements in code reading: Relaxing the linear order." In *Proc. 2015 IEEE 23rd Int. Conf. Program Comprehension (ICPC 2015)*, Florence, Italy; 2015: 255-265. 10.1109/ICPC.2015.36.
21. Körner C. Sequential processing in comprehension of hierarchical graphs. *Appl Cogn Psychol*. 2004;18(4):467-480. doi:10.1002/acp.997
22. Huang W. Establishing aesthetics based on human graph reading behavior: two eye tracking studies. *Pers Ubiquitous Comput*. 2013;17(1):93-105. doi:10.1007/s00779-011-0473-2
23. Pohl M, Schmitt M, Diehl S. "Comparing the readability of graph layouts using eyetracking and task-oriented analysis." In *Proc. 5th Eurographics Conf. Computational Aesthetics in Graphics, Visualization and Imaging (Computational Aesthetics 2009)*, Victoria, Canada; 2009: 49-56. 10.5555/2381286.2381296.
24. Sharif B, Maletic J. "The effect of layout on the comprehension of UML class diagrams: A controlled experiment." In *Proc. 2009 5th IEEE Int. Workshop on Visualizing Software for Understanding and Analysis (VISSOFT 2009), Edmonton, Canada*; 2009: 11-18. 10.1109/VISSOF.2009.5336430.
25. Hu X, Bačić D. Exploratory discoveries from eye-tracking tests of wertheimer's gestalt patterns. *Leonardo*. 2021;54(5):517-523. doi:10.1162/leon_a_02005
26. Behe BK, Campbell BL, Khachatryan H, et al. Incorporating eye tracking technology and con-joint analysis to better understand the green industry consumer. *HortScience*. 2014;49(12):1550-1557. doi:10.21273/HORTSCI.49.12.1550
27. *R: A Language and Environment for Statistical Computing*. 2022. [Online]. Available: https://www.R-project.org/
28. *Rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. 2021. [Online]. Available: https://CRAN.R-project.org/package=rstatix
29. Field A, Hole G. *How to design and report experiments*. SAGE Publications; 2002.
30. Albers C, Lakens D. When power analyses based on pilot data are biased: inaccurate effect size estimators and follow-up bias. *J Exp Soc Psychol*. 2018;74(1):187-195. doi:10.1016/j.jesp.2017.09.004
31. Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends Sport Sci*. 2014;21(1):19-25.

How to cite this article: Grabinger L, Hauser F, Mottok J. On the perception of graph layouts. *J Softw Evol Proc*. 2023;e2599. doi:10.1002/smr.2599