Diagram Size vs. Layout Flaws: Understanding Quality Factors of UML Diagrams
Understanding Quality Factors of UML Diagrams

Störrle, Harald

Published in:
ESEM '16 Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement

Link to article, DOI:
10.1145/2961111.2962609

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Störrle, H. (2016). Diagram Size vs. Layout Flaws: Understanding Quality Factors of UML Diagrams: Understanding Quality Factors of UML Diagrams. In *ESEM '16 Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (pp. a31). Association for Computing Machinery. https://doi.org/10.1145/2961111.2962609

Diagram Size vs. Layout Flaws: Understanding Quality Factors of UML Diagrams

Harald Störrle
Department of Applied Mathematics and Computer Science
Technical University of Denmark, Matematiktorvet, 2800 Lyngby, Denmark

ABSTRACT

CONTEXT: Previously, we have defined the notion of diagram size and studied its impact on the understanding of UML diagrams. Subsequently, questions have been raised regarding the reliability and generality of our findings. Also, new questions arose regarding how the *quality* of diagrams could be defined, and how it interacts with diagram size.
GOAL: We pursue three goals. First, we want to increase the validity of our research by analyzing a substantially larger data set than before. Second, we broaden the generalizability of our results by including two more diagram types. Our main contribution, though, is our third goal of extending our analysis aspects of diagram quality.
METHOD: We improve our definition of diagram size and add a (provisional) definition of diagram quality as the number of topographic layout flaws. We apply these metrics on 60 diagrams of the five most commonly used types of UML diagram. We carefully analyze the structure of our diagram samples to ensure representativeness. We correlate diagram size and layout quality with modeler performance data obtained in previous experiments. The data set is the largest of its kind ($n = 156$).
RESULTS: We replicate earlier findings, and extend them to two new diagram types. We provide an improved definition of diagram size, and provide a definition of topographic layout quality, which is one more step towards a comprehensive definition of diagram quality as such. Both metrics are shown to be objectively applicable. We quantify the impact of diagram size and quality on diagram understanding.
CONCLUSIONS: The overall results of previous studies are confirmed, while our previous recommendations for creating better diagrams are revised and refined.

1. INTRODUCTION

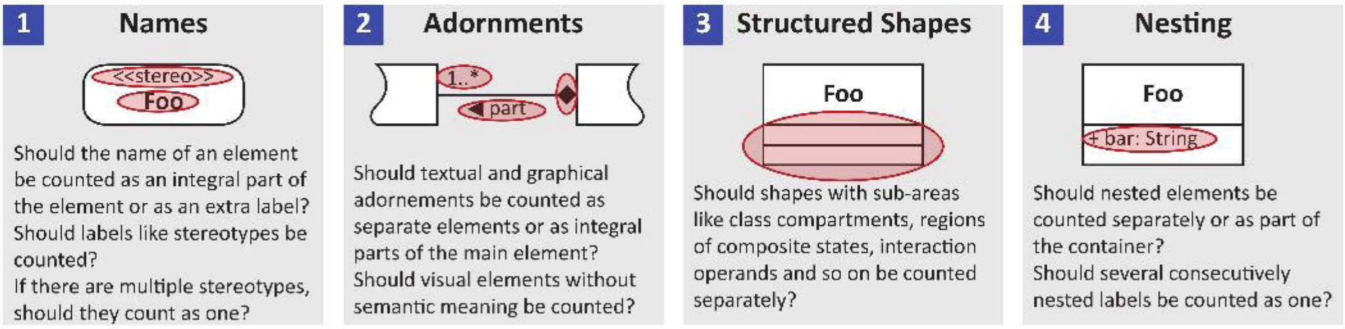
The Unified Modeling Language (UML) has been the "*lingua franca of software engineering*" for well over a decade. It is a generally held belief that visual languages are superior to textual languages in that they support human perceptual and thought processes, and that this is also true for the UML, in fact, that this is a major reason for the success of UML. However, there are actually few research results to support this belief. There *is* a large body of experimental results on the layout of UML class diagrams and how it affects human understanding and problem solving, but the findings are ambiguous, and sometimes unintuitive. In particular, only very small effects have been found in vitro. For instance, Eichelberger and Schmid note that "*We could not identify [...] a significant impact [by diagram quality].*" (cf. [9, p. 1696]).

On the other hand, practical experience in industrial software projects suggests a much higher impact of good or bad layout. In particular, our initial hypothesis is that with increasing size and decreasing quality, modeler performance in model understanding tasks decreases. This has indeed been supported by our earlier work (see [24, 25]). Closer inspection of our data suggested, however, that the size of the models visualized in the diagrams might be a relevant factor. In [26], we have explored notions of diagram size and re-examined existing sets of experimental data. We found that increasing diagram size correlates to decreasing model understanding performance of modelers. We also conjectured diagram layout quality matters more with increasing diagram size: small diagrams are easy to use irrespective of the layout quality: modelers simply cope with bad layout. With increasing diagram size, however, the visual and/or mental capacity of a modeler is stretched, so that the layout quality reduces modeler performance. In other words, layout quality matters more, and is more apparent for larger diagrams. Based on our findings we derived a recommendation for a limit of diagram size which is helpful as a guideline to inexperienced modelers, such as students.

This previous work has raised a number of questions. Some have questioned the validity of our study, pertaining mostly to the number of diagrams used. Others have questioned the generalizability to other diagram types, suggesting different diagram types have very different characteristics, resulting in different size limits. We address these concerns by doubling the data set used in the present study to almost 14,000 data points (1,207 experimental items) from 156 participants, and adding two more UML diagram types so that our results now reflect the five most commonly used UML diagrams [14, 3].

A second class of questions centered on the notions of size and quality, suggesting that diagram quality should be quantified, too. By definition, quality is difficult to quantify. However, there are *some* aspects that are fairly straightforward. For instance, line crossings or bends clearly are

ESEM '16, September 08 - 09, 2016, Ciudad Real, Spain
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4427-2/16/09...$15.00
DOI: http://dx.doi.org/10.1145/2961111.2962609



Should the name of an element be counted as an integral part of the element or as an extra label? Should labels like stereotypes be counted? If there are multiple stereotypes, should they count as one?

Should textual and graphical adornements be counted as separate elements or as integral parts of the main element? Should visual elements without semantic meaning be counted?

Should shapes with sub-areas like class compartments, regions of composite states, interaction operands and so on be counted separately?

Should nested elements be counted separately or as part of the container? Should several consecutively nested labels be counted as one?

Figure 1: Decision points in counting rules for diagram size.

diagram layout flaws and ought to be avoided. Thus, we define the following research questions for our paper.

- **RQ 1:** Do the results of our previous study hold up when analyzing a (much) larger, more diverse data set?
- **RQ 2:** Diagram quality is an elusive notion: can we characterize it in a precise way, at least part of it?
- **RQ 3:** What is the impact of diagram quality on model understanding, as compared to diagram size?

In answering these questions, we aspire to expand our understanding of the factors responsible for the understandability of models, work towards a comprehensive definition of diagram layout quality, and provide practical guidelines.

2. SIZE OF UML DIAGRAMS

In [26], we have defined *"diagram elements"* as any line segment, shape, or textual label that appears in a diagram; we proposed to use the number of diagram elements as the size of a diagram. Other, more "intuitive" metrics we have considered in [26] added considerable complexity but were still highly correlated, so we discarded them again. In [26], we have defined *"diagram elements"* as any line segment, shape, or textual label that appears in a diagram; we proposed to use the number of diagram elements as the size of a diagram. Other, more "intuitive" metrics we have considered in [26] added considerable complexity but were still highly correlated, so we discarded them again.

However, this definition has two shortcomings. First, by counting line *segments* rather than lines, line bends contribute to diagram size although they represent a *quality* aspect [9]. Thus, our metric mixed aspects of diagram size and quality, impeding the individual analysis of these two factors. Second, clear and straightforward as our definition may be, assessing diagram size by different people yielded different results. We discovered four ambiguities (see Fig. 1).

1. **Names:** should element names be counted as labels, or are they integral parts of named element? Should attachments like stereotypes count as separate labels?
2. **Adornments:** Should textual and graphical adornments such as multiplicities and arrow heads be counted as separate labels and shapes, respectively, or should they be considered part of the adorned element? What about visual elements without semantic counterpart, such as triangles indicating reading direction of associations: should they be counted as visual elements?
3. **Structured Shapes:** Should structured shapes be counted as a single shape, or should all the sub-structures be counted by themselves? This applies to classes with several compartments, regions of concurrent composite states, operands of interaction fragments with binary operators, but also to swim-lanes in activity diagrams.
4. **Nesting:** If sub-elements are nested within a simple or structured element, should they be counted separately? Should every line be counted as a single label, even if it is a continuous sentence in a comment, or should consecutive lines be counted as one label?

In order to resolve these cases, we offer the following refined definition for diagram element.

Definition 1 *A* ***diagram element*** *is any line, shape, or textual label that appears in a diagram and*

(a) can be positioned within the diagram by itself, or

(b) can be shown or hidden by itself, or

(c) contains other diagram elements.

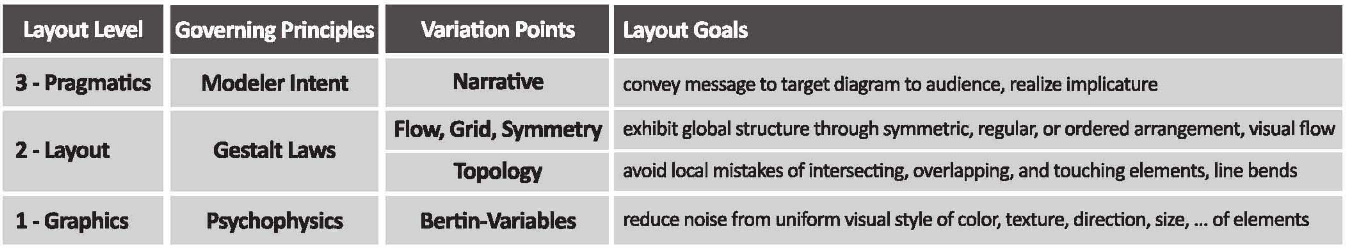
The ***size of a diagram*** *is the number of its diagram elements.*

Applying this definition to the above questions yields this.

1. Names can be neither hidden nor moved so they do not count as separate elements. Stereotypes, on the other hand, can be hidden so they do count as labels.
2. Adornments with fixed position relative to the adorned element are not counted, e.g., arrow heads, and aggregation-diamonds. Adornments that *can* be moved include multiplicities, association names, and transition guards.
3. Class compartments can be hidden individually, so they count as extra elements, unless they are empty.
4. Nested elements are counted because they can be hidden and ordered in most tools.

Observe that we make no reference to the *model* presented by a diagram other than whether diagram elements refer to *separate* model elements or not. Thus, the rules apply irrespective of whether visual elements do or do not have a semantic counterpart. Some tools allow to collapse substructures, thus hiding some diagram elements. In our definition, this corresponds to different diagrams.

In the process of teaching modeling, we are faced with UML models of all kinds and qualities at a rate of several hundred (sic) per year. Using them as test cases for our metric definition, we found our metric to be simple enough



Layout Level	Governing Principles	Variation Points	Layout Goals
3 - Pragmatics	Modeler Intent	Narrative	convey message to target diagram to audience, realize implicature
2 - Layout	Gestalt Laws	Flow, Grid, Symmetry	exhibit global structure through symmetric, regular, or ordered arrangement, visual flow
		Topology	avoid local mistakes of intersecting, overlapping, and touching elements, line bends
1 - Graphics	Psychophysics	Bertin-Variables	reduce noise from uniform visual style of color, texture, direction, size, ... of elements

Figure 2: There are three layout levels of diagrams. In this paper, we are solely concerned with the topological layout (level 2a).

to be readily understood by students. Also, the rule set is consistent and covers all of UML.

The examples in Fig. 1 focus mainly on class diagrams because this is where most of the problems arise: The counting rules apply equally to all of UML. It remains to be seen, however, whether it is sufficient to cover visual modeling languages other than UML. Fig. 4 below shows an example of applying these counting rules, and contrasts it with the results yielded by the rules defined in [26].

3. QUALITY OF UML DIAGRAMS

Based on the notion of diagram size, we now proceed to the notion of quality. There are three dimensions to the design of diagrams that affect its quality: the graphic level, the layout level, and the pragmatic level.

1. **Graphics** refers to basic perceptual features as studied by perceptual psychology [29, 28]. Here, we are concerned with graphical properties like color, line thickness, texture, shape, and so on.
2. **Layout** refers to all aspects of arranging elements of a diagram. This can be subdivided into local and global aspects that focus on topological features and questions of flow and symmetry, respectively, that are governed by the laws of Gestalt psychology. Most of the empirical research on UML diagrams focuses on topological aspects, e.g., [21, 7, 10, 30, 18].
3. **Pragmatics** refers to the value of diagrams as a communication medium. This is governed by the modeler intent, narrative to be conveyed, medium constraints and affordances, and target audience (see [15, 12]).

Generally, higher level concerns may take precedence over lower level concerns when it comes to creating "good" diagrams. For instance, in order to highlight a certain diagram element to the audience, it is quite effective to violate graphical uniformity and highlight it in a contrast color. Or, the modeler may choose to presented a diagram element in a way that breaks the symmetry or flow of the overall layout. Similarly, in order to achieve a good overall layout, topological flaws like the occasional line crossing may be accepted. Clearly, such trade-offs are difficult to make, let alone to automate. But even seemingly simple aspects of diagram topology offer more complexity than meets the eye.

Previous research on general graphs [2, 16] as well as on UML (class) diagrams [21, 18, 17, 10, 1, 7, 30, 9] has studied layout aspects of diagrams, in particular intersecting, touching, and overlapping elements, line bends, and redundant lines. There is clear evidence that they negatively affect the understandability of a diagram and should be avoided. While [16, 9] also discuss higher-level layout aspects like symmetry and flow, there is much less agreement and empirical evidence for them than for the low-level layout aspects. So, we make this our starting point and consider all the low-level topographic problems listed in [9, pp. 1689] as "diagram flaws". However, as with diagram size, what appears to be a straightforward definition becomes difficult when operationalizing it. Consider the following ambiguities.

- Bends should be considered flaws, but what about curves? Should the opening angle be considered, as [16] suggests?
- Also, it is often recommended to merge lines "where appropriate", but exactly when is that the case?
- Probably the biggest issue are the many forms of intersections, including line crossings, and obscuring/touching elements. Which of these should be considered as flaws, for instance, should we count a line crossing that is mitigated by a "bump" as a flaw at all? Should the crossing angle be taken into consideration [7, p. 65]? Should intersecting sub-elements be counted extra? At what distance are two elements considered as touching each other? What about unavoidably line crossings, or elements that are overlapping because that is an expressive element of the visual language in question?

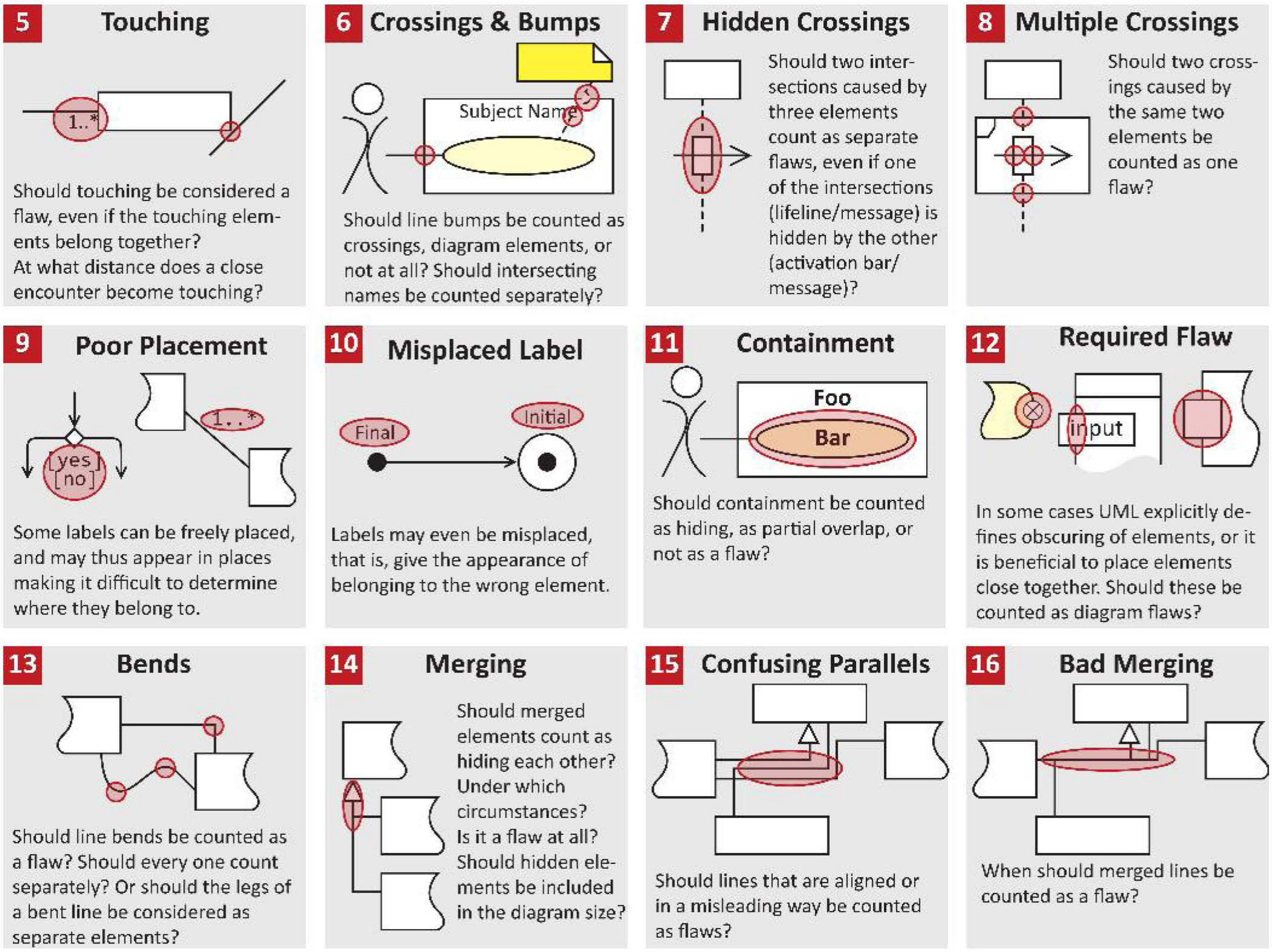
A complete list of problems is shown in Fig. 3. We decide these issues such that whatever implies that a modeler has to take a decision is considered a layout flaw. Typical examples are poor placement (case 9 in Fig. 3) and confusing parallels (case 15). Conversely, a line crossing is *not* counted, if it is invisible (case 7). Similarly, when two elements overlap because the language syntax demands that they do, we do not consider this a flaw (case 12). For simplicity, we do not distinguish between degrees of flaws, such as the degree of opening of a bend, a case considered by Purchase in [16].

We also posit, that the list of problems presented in Fig. 3 is exhaustive, that is, the cases defined by these rules constitute *all* flaws at the level of diagram topology in the sense of Fig. 2. In the present paper, we focus exclusively on diagram flaws on this level in order to allow a comprehensive treatment. We yield the following definition of (topological) diagram flaws and (topological) quality of a diagram.

Definition 2 *A* ***diagram flaw*** *is an instance of*

(a) Bends of lines are considered flaws.

(b) Intersections are considered flaws if they are visible and not a syntactic element of the language.



Should touching be considered a flaw, even if the touching elements belong together? At what distance does a close encounter become touching?

Should line bumps be counted as crossings, diagram elements, or not at all? Should intersecting names be counted separately?

Should two intersections caused by three elements count as separate flaws, even if one of the intersections (lifeline/message) is hidden by the other (activation bar/ message)?

Should two crossings caused by the same two elements be counted as one flaw?

Some labels can be freely placed, and may thus appear in places making it difficult to determine where they belong to.

Labels may even be misplaced, that is, give the appearance of belonging to the wrong element.

Should containment be counted as hiding, as partial overlap, or not as a flaw?

In some cases UML explicitly defines obscuring of elements, or it is beneficial to place elements close together. Should these be counted as diagram flaws?

Should line bends be counted as a flaw? Should every one count separately? Or should the legs of a bent line be considered as separate elements?

Should merged elements count as hiding each other? Under which circumstances? Is it a flaw at all? Should hidden elements be included in the diagram size?

Should lines that are aligned or in a misleading way be counted as flaws?

When should merged lines be counted as a flaw?

Figure 3: Decision points in counting rules for diagram flaws.

(c) *Touching elements are considered a flaw, unless they have close syntactic or semantic association.*

(d) *Sets of merged lines or aligned lines that are close together are considered a flaw, unless they have the same type and share exactly one of their endpoints.*

(e) *Two flaws are fused into one flaw, if they are very close together and caused by the same intersecting elements.*

*The **topological quality** of a diagram is defined as the number of flaws it contains.*

Clearly, we need make more precise the notions of "close" and "very close together". Based on human physiology, we argue that this should be the case for any two elements that are less than 5*mm* and .5mm apart, respectively: for detailed visual perception (particularly reading), humans use the receptors placed in a particular structure of the retinal surface, called the fovea centralis. This area corresponds to approximately 5° of the human visual field. Assuming the diagram is displayed on a laptop or desktop screen and the modeler is reading the screen at the ergonomically recommended reading distance (about 50cm). Then the diagram area corresponding to the area of the fovea centralis is a circle with approx. 22mm radius. We interpret "close" and "very close" as 10% and 1% of the diameter, respectively, which results in distances on the diagram of approximately 5*mm* and 0.5*mm*, respectively.

4. UNDERLYING STUDIES

We have previously presented two studies about the impact of layout quality to model understanding performance [24, 25]. In this paper we progress towards formal definitions of the notions of diagram size and quality, and validate them using the data obtained previously. We restrict ourselves to the aspects required for the given context and refer the reader to the original publications for more detail.

4.1 Study Design

This paper does report new primary studies, but re-analyzes existing data from two previous studies [24, 25]. Nevertheless, we have to discuss the study design used in the primary studies to allow the reader to assess the data we present.

In [24, 25] we report studies that data of which are re-analyzed in the current paper. Both studies consisted in three similar experiments on different populations of CS students. Participants were randomly assigned to one of four different sequences of nine tasks presented by paper questionnaires,

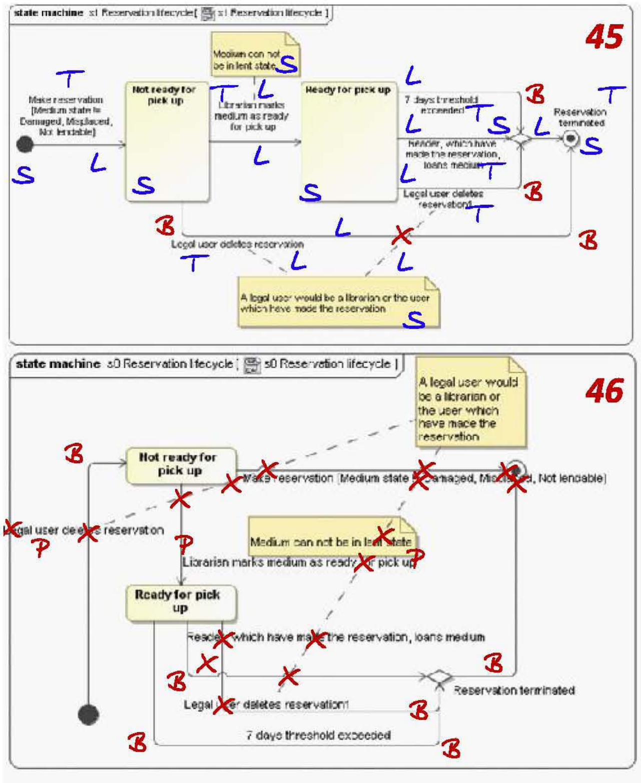

state machine s1 Reservation lifecycle[s1 Reservation lifecycle]
45
Medium can not be in lent state
Make reservation [Medium state != Damaged, Misplaced, Not lendable]
Not ready for pick up
Librarian marks medium as ready for pick up
Ready for pick up
7 days threshold exceeded
Reader, which have made the reservation, loans medium
Legal user deletes reservation
Reservation terminated
Legal user deletes reservation
A legal user would be a librarian or the user which have made the reservation
state machine s0 Reservation lifecycle [s0 Reservation lifecycle]
46
A legal user would be a librarian or the user which have mace the reservation
Not ready for pick up
Make reservation [Medium state != Damaged, Misplaced, Not lendable]
Legal user deletes reservation
Medium can not be in lent state
Librarian marks medium as ready for pick up
Ready for pick up
Reader, which have made the reservation, loans medium
Reservation terminated
Legal user deletes reservation
7 days threshold exceeded
T
L
S
B
X
P

Diagram	45/46
S Shape	7
T Text	7
L Line	10
Elements	24

Diagram	45	46
X Intersection	1	16
B Bend	4	6
P Positioning	0	3
A Alignment	0	0
Flaws	5	25

Diagram	45	46
Flaws/Elements	0.21	1.04

Figure 4: Examples for measuring diagram size and topological quality of diagrams for different layouts of the same model. The blue and red characters on the diagrams indicate diagram elements and flaws by type. Both diagrams have the same size, by definition.

plus demographic questions. Each task consisted of one page showing a UML diagram, and ten questions about the model presented in the diagram. The diagram sequences were balanced wrt. type, size, and layout quality. The first study [24] contained class, activity, and use case diagrams, the second one [25] replaced activity and use case diagrams by state machine and sequence diagrams, respectively. There were twelve diagrams for each diagram type.

The dependent variables included score of the comprehension questions. We also asked preference questions and recorded task completion duration, but this data is not analyzed in this paper, and thus subsequently neglected. The independent variables were the experience level of the participants, the diagram type, and the diagram size and layout quality. Between them, the six experiments conducted in the two studies presented findings based on 60 diagrams of five types, conducted on five different populations with a total of 156 participants (completion rate over 80%). This is, by far, the largest data set of its kind, in any of these dimension.

4.2 Diagram samples

The complete set of all diagrams contained in our study, as well the questionnaires, and the raw data are available for download at http://bit.ly/1RJrv8K. An online version of the complete experiment is also publicly available at http://goo.gl/forms/6cKjvdhzwp.

The diagrams used in the experiments have been created by students as part of a their course work on a Requirements Engineering class with approximately 500-700 person-hours of effort per model.[1] From these case studies, we selected typical diagrams and prepared a second version of the diagram, where we improved the layout as much as we could.

Clearly, there is a danger that the diagrams in our sample are biased and might thus influence the findings we base on them. We argue, however, that they are *representative* of UML diagrams in general. In order to support this claim, we have analyzed our sample in several ways to see whether there is a bias with regards to size or number of flaws.

Fig. 5 shows the size and quality of the diagrams in our sample in terms of diagram elements and number of layout flaws. The counts of elements and flaws are stacked to highlight the difference between the current definition of diagram size and the earlier one proposed in [26], which counted (some flaws) as part of diagram size as discussed in Section Section 2. It is obvious that for each diagram type, there is a spread of sizes and qualities, so there is no bias within each group. Clearly, there are differences across groups, though: Use Case diagrams tend to be somewhat smaller than other diagrams, and class diagrams sometimes get fairly large. However, this is indeed typical of practical models in industry, as the author can assert from many years of industrial practice. For reference, we also compared our sample with [13], the largest publicly available class diagram repository. As of March 10th, 2016, the Gothenburg repository contained 810 class diagrams with an average number of 35.4 elements in them (min: 1, max: 282). The distribution is shown in Fig. 5 with index G, it is obviously in the range of the sizes in our samples. In other words, our sample is indeed representative, as far as size is concerned.

As far as layout quality is concerned, looking at the number of flaws is somewhat misleading, as large diagrams naturally allow many more flaws than small diagrams. So, we consider the flaw rate instead of the flaw number (i.e., $\frac{flaws}{elements}$). Fig. 6 shows the distribution of the flaw rate in our sample. Clearly, there is an even distribution over all degrees of layout quality, both regarding the overall sample, and for each diagram type. The typical size differences between different diagram types that we have noticed before are visible also in the flaw rates. From this analysis we conclude that the diagrams in our sample cover the whole spectrum of sizes and qualities, for each diagram type considered. In this sense, our sample is representative of the spread that exists in the true population of diagrams.

4.3 Study participants

The participants of our studies are students in various

[1] The course is worth 10 ECTS points corresponding to 280h of work. Teams of 4-7 students collaborate over a period of twelve weeks, most of the time in a very practical setting. Combining these facts with the (conservative) assumption that half the effort is spent on modeling, and further assuming that, on average, students do as much work as they are supposed to, this amounts to 560-680h of effort per model.

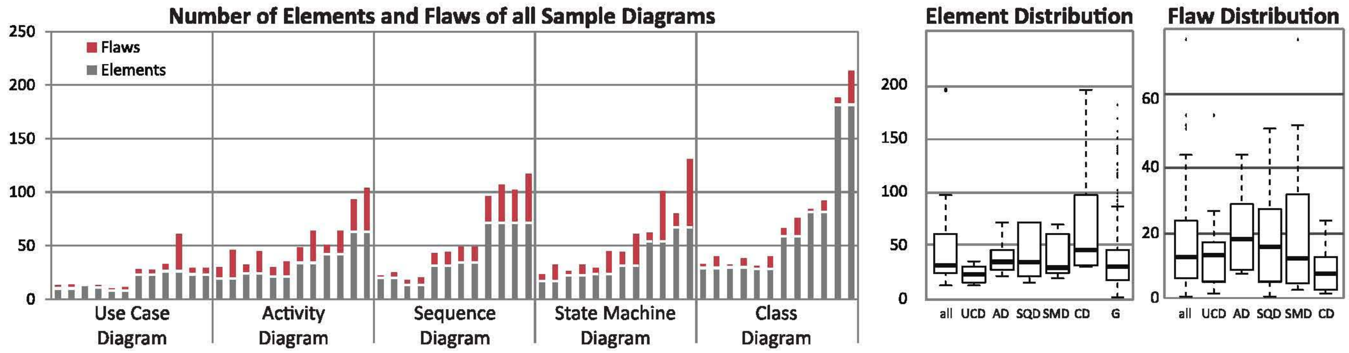


Figure 5: Distribution of diagram sizes per diagram type: the bottom/grey bars show numbers of elements, the top/red bars show number of layout flaws per diagram. The boxplots to the right show distribution of elements and flaws, respectively, in total and by diagram. The box with index G refers to [13].

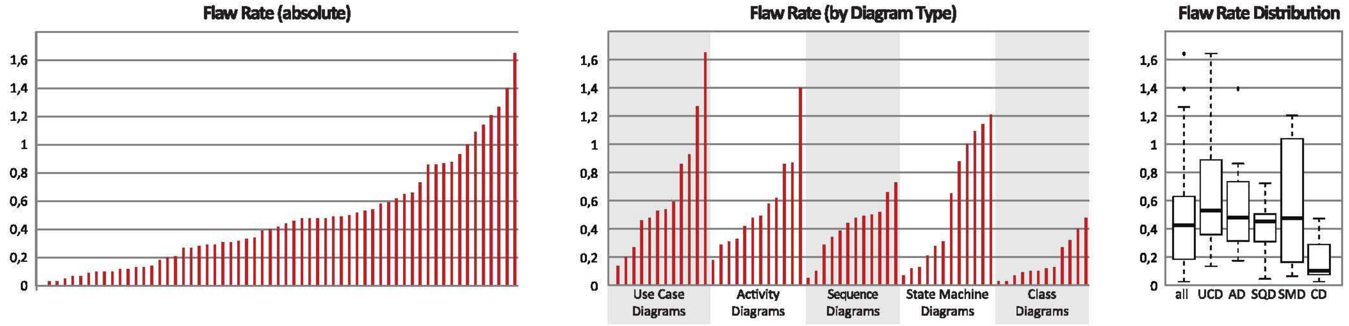


Figure 6: Distribution of layout flaws overall (left), by diagram (middle), and in summary (right). In the graphs on the left and in the middle, each bar represents one sample diagram.

computer science programs at the Technical University of Denmark in Lyngby and the University of Augsburg (see Table 1 for an overview). One may argue that this population does not represent the true population of modelers, which consists of practitioners with substantial professional experience. However, between a third and half of our students have part-time programming jobs in industry, and are about to become professionals immediately after completing their degree. In that sense the study participants are fairly representative of junior developers. Likely, more senior developers will have a greater level of expertise which will result in better performance in the tasks tested (see our analysis of expertise levels in [25]). On the other hand, professionals with a technical background are not the only ones to use models, and it is fair to expect lower expertise levels for this audience, which constitutes the *opposite* bias.

Additionally, observe that the population we have tested is unusually large for experiments of this kind: many classic psychological experiments are conducted with populations a fraction of this size (cf. [11, p. 56]: *"it should be remembered that an N of 25 is a good deal larger than the numbers sometimes reported!"*). So, there is no reason to assume that the population tested in our studies are distorting the results in any particularly way. In fact, we should assume a much smaller degree of variation than in many existing experiments.

4.4 Threats to validity

External validity The selection of the models and diagrams may be a source of bias. However, we applied objective and rational criteria to the selection. Compared to the related work, we used more diagram types (three rather than just one or two), more models, and more realistic models. The layouts for the models were, to a large degree, used-as-found, that is, they were created under realistic conditions by people unconnected to this study. Additionally, our study is based on a comparatively large number of participants. Therefore, the present study can be exhibits a much larger degree of validity than previous work. We expect our results to hold for UML models *in general*, i.e., we expect a markedly higher degree of external validity than previous contributions in this field.

Table 1: Demographic data on the participants of all experiments, "completion" refers to the completion rate on core questions.

Experiment	male	female	all	completion
A (BSc)	23	4	27	90.3%
B (BSc)	21	1	22	86.6%
C (MSc)	27	2	29	67.6%
D (BEng)	29	3	33	75.1%
E (MSc)	29	5	34	82.6%
F (Elite)	10	1	11	90.1%
all	**139**	**17**	**156**	**82.1%**

Internal validity Great care has been taken to provide systematic permutations of diagrams and question sequences to avoid carry-over effects ("learning"). Any such effects would occur similarly for all treatments and, thus, cancel

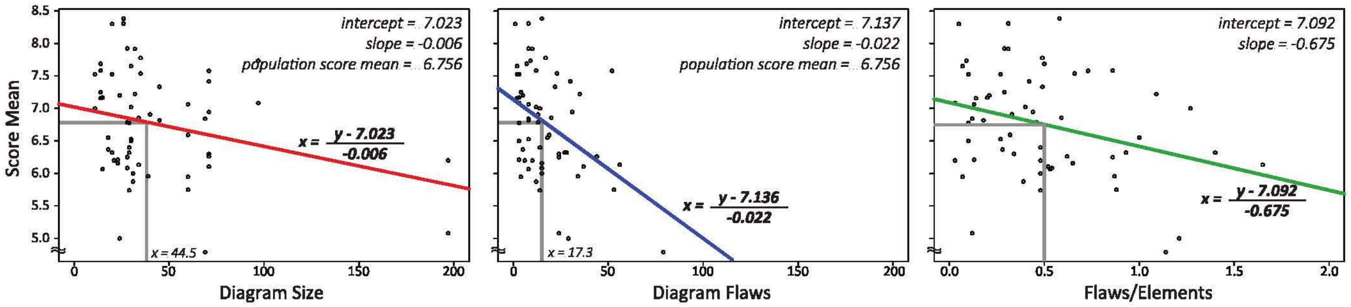


Figure 7: Modeler performance as mean understanding task score vs. diagram size (left), number of flaws (middle), and topological quality (right). The trend-lines are created from linear models. The two outliers in the left diagram have almost no impact on the slope.

each other out. Participants have been assigned to tasks randomly. We can also exclude bias through the experimenter himself, since there were only written instructions that apply to all conditions identically. We correlated it with different measures, each of which was measured in multiple different ways to reduce the danger of introducing bias through the experimental procedure.

Construct validity We have previously argued for the validity of the element count as a size metric [26]. Clearly, the number of flaws is part of the quality of diagrams, though there are likely other factors as well—in particular those that we have described in Section 3. What the relative magnitude of these factors will have to be answered by future research.

Conclusion validity We have consistently provided statistical significance levels and effect sizes (using Cohen's convention). While many of the results are not significant due to the relatively small number of diagrams per type, almost all results are consistently pointing in the same direction. We assume a linear correlation between variables prima facie, but this is justified by an earlier ANOVA-analysis where the squared terms were much too small to have a significant impact.

5. RESULTS

5.1 Validation of counting rules

The first step in our validation is to ensure the counting rules defined above are clear enough to be applied by different people. To that end, we have asked two junior colleagues to count the same test suite of 60 diagrams, instructed only by the counting rules described above. We compared the results and discussed deviations, which resulted in no refinement of the rules. The ratings show a very high correlation using Pearson's r (Cohen's κ applies only to categorical data), see Table 2. This means that the operationalization of the counting rules is sufficiently clear to yield reliable metrics results across raters.

Table 2: High inter-rater correlation of manual counting indicates unambiguous counting rules.

	Elements	Flaws	Flaw Rate
r	0.994	0.977	0.979
p	$< .10^{-15}$	$< .10^{-15}$	$< .10^{-15}$
sig	***	***	***

5.2 Size and quality vs. modeler performance

As outlined above, our initial hypothesis is that with increasing size and decreasing quality, modeler performance in understanding tasks decreases. Plotting the diagram size and quality as defined above against the understanding performance on all diagrams yielded the scatter plots shown in Fig. 7. The trend-lines represent fitted linear models. As expected, the mean score decreases when the number of diagram elements and flaws increases.

We then tested computed the correlations of the data split into various subgroups (see Table 3). Correlations were calculated using Pearson's product-moment correlation. Following Cohen's convention, we assess the effect size of a correlation of up to 0.3 to as small (S), as large (L) for values over 0.4, and as medium (M) for values in between. When looking at all diagrams and all participants, respectively, almost all correlations are statistically significant, both for score mean and score variance. Splitting up the data for individual diagram types or sub-populations, most correlations are not significant any more due to the reduced sample size. Where the sub-samples are large (e.g., for class diagrams), we still see significant correlations, and for smaller sub-groups, almost all correlations are consistent. As we have noticed in [26], the populations with higher capabilities are much less affected by large size and poor diagram quality.

Most results are not statistically significance. It is remarkable, though, that they consistently point in the same direction, indicating that decreasing size and number or rate of flaws correlate to better performance. The same is found when splitting the correlations by expertise level, which also yields much higher significance. This indicates that the variation of the results is impacted more through expertise than through diagram size and quality, which is consistent with our previous findings [26].

5.3 Optimal diagram size

In [26] we have used the correlation data to derive a recommendation for optimal diagram size. Fitting a linear model to the correlation data (see Fig. 7) we obtained an intercept of 7.137 and a slope of −0.022. Inserting the population mean score of ca. 6, we computed the "center size", which we define as the number of diagram elements for which most modelers should be able to perform best on many diagrams. For the given values, the center size is approximately 50. Two questions immediately arise for the new analysis presented in this paper: (1) does changing the counting rules have an impact and if so, which, and (2) does "optimal size"

Table 3: Pearson's product-moment correlation between diagram size and quality, and modeler performance measured as mean and variance of objective performance in understanding tasks: *r* is Pearson's *r*, *ES* is effect size in Cohen's classification, followed by the *p*-value and its significance level.

	Diagram Size				Diagram Flaws				Diagram Flaw Rate			
Score Mean	*r*	ES	*p*	SIG	*r*	ES	*p*	SIG	*r*	ES	*p*	SIG
All Diagrams	−0.270	M	0.037	*	−0.419	L	< .001	***	−0.312	M	0.015	*
Use Case	−0.634	L	0.027	*	−0.464	L	0.128		−0.256	M	0.422	
Activity	−0.115	S	0.722		−0.523	L	0.081	.	−0.514	L	0.088	.
Sequence	−0.387	M	0.215		−0.237	M	0.458		−0.179	S	0.577	
State Machine	−0.305	M	0.335		−0.578	L	0.049	*	−0.466	L	0.127	
Class	−0.539	L	0.071	.	−0.616	L	0.033	*	−0.121	S	0.707	
Score Variance	*r*	ES	*p*	SIG	*r*	ES	*p*	SIG	*r*	ES	*p*	SIG
All Diagrams	0.392	M	0.002	*	0.262	M	0.043	*	0.189	S	0.149	
Use Case	0.492	L	0.104	.	0.606	L	0.037	*	0.578	L	0.049	*
Activity	0.117	S	0.718		0.260	M	0.414		0.241	M	0.450	
Sequence	0.288	M	0.364		0.065	XS	0.840		−0.235	M	0.462	
State Machine	0.093	S	0.773		0.236	M	0.459		0.211	M	0.510	
Class	0.833	XL	< .001	***	0.227	M	0.477		−0.328	M	0.297	
Score Mean	*r*	ES	*p*	SIG	*r*	ES	*p*	SIG	*r*	ES	*p*	SIG
All Participants	−0.270	M	0.037	*	-0.419	L	< .001	***	−0.312	M	0.015	*
BEng	−0.342	M	0.044	*	-0.510	L	0.002	**	−0.489	L	0.003	**
BSc	−0.196	S	0.251		-0.364	M	0.029	*	−0.330	M	0.049	*
MSc	−0.223	M	0.087	.	-0.276	M	0.033	*	−0.165	S	0.207	
Elite	−0.070	S	0.687		-0.173	S	0.312		−0.155	S	0.366	
Score Variance	*r*	ES	*p*	SIG	*r*	ES	*p*	SIG	*r*	ES	*p*	SIG
All Participants	0.392	M	0.002	**	0.262	M	0.043	*	0.189	M	0.149	
BEng	0.197	M	0.264		−0.047	S	0.792		−0.203	M	0.250	
BSc	0.349	M	0.037	*	0.190	M	0.268		0.213	M	0.213	
MSc	0.251	M	0.053	.	0.212	M	0.104		0.143	S	0.276	
Elite	−0.006	XS	0.975		−0.121	S	0.488		−0.089	XS	0.613	

recommendation still hold with the augmented data set?

To answer these questions we computed linear models for both data sets separately and combined, see Table 4: each graph defines its intercept and slope. To better account for population variance, we used the mean of the respective population rather than the median. Regarding the first question we compare the score center sizes for experiments D-F with the ones reported in [26]. By applying the new counting rules (and using the mean instead of the median), the center size increases from 50 to 65. This demonstrates that the new metric defined in this paper excludes some phenomena that negatively impact modeler performance—which is exactly what we should expect, as the previous metric implicitly included some diagram types of flaw.

As for the second question, comparing the center sizes yielded from the two sets of experiments shows an even larger difference, but the result obtained for the overall data set yields almost the same result as for the data from experiments A-C. This is likely the case because the population of the first experiment was more homogeneous in terms of capability; the populations in the second three experiments were selected specifically to have a wide range in expertise level.

5.4 Optimal diagram quality

We now turn to the question of diagram quality. As before, we consider not just the number of flaws, but also the flaw rate in order to have a measure that corrects for diagram size. It is obvious that the impact of the number of flaws is much greater than the number of elements. Unlike the results for diagram size, the intercepts, slopes, and center sizes are almost identical, across data sets. That means also, that this factor has an impact that is much less affected by expertise or diagram type, which might mean that it is affecting a different, more basic cognitive mechanism than the one dealing with diagram size. We translate these findings into the following guidelines.

- The guidelines for diagram size proposed in [26] hold.
- The flaw rate of a diagram should not exceed 0.5.
- The number of flaws should not exceed 15-20.
- Improving diagrams should prioritize reducing the number of flaws over reducing the number of elements.

6. RELATED WORK

The main focus of previous work on UML diagram types and their layout has been with one of four aspects: diagram comprehension (cf. [22, 19] and/or user preference (cf. [18, 27]), automatic layout (cf. [7, 10, 16, 8, 4]), or one of a variety of diagram inference tasks, e.g., program understanding based on visualizations (cf. [29]), or the role of design patterns in understanding (cf. [22, 23]).

Table 4: Recommendations for diagram size and quality based on population mean and linear regression.

	Experiments A-C & D-F	**Experiments A-C**	**Experiments D-F**
	population score $\mu = 6.756$	**population score $\mu = 6.737$**	**population score $\mu = 6.755$**
Diagram Size	$7.023/-0.006:$ **44.2**	$7.019/-0.005:$ **48.5**	$6.865/-0.002:$ **64.6**
Diagram Flaws	$7.137/-0.022:$ **17.7**	$7.118/-0.023:$ **16.0**	$7.150/-0.024:$ **16.5**
Diagram Flaw Rate	$7.137/-0.022:$ **0.50**	$7.118/-0.023:$ **0.51**	$7.150/-0.024:$ **0.54**

The layout of graphs (in the mathematical sense) has been a longstanding research challenge, both with respect to automatic layout and to various aspects of usability, e.g., diagram comprehension, user preferences, and diagrammatic inference. Based on the rich knowledge on general graphs, research on the layout of UML has started with those of UML's notations that are closest to graphs, namely, class diagrams (cf. [21, 7, 10, 30, 18]), and, to a lesser extent, communication diagrams (see e.g. [17] who use UML 1 terminology). Other types of UML diagrams, in contrast, have only attracted little interest so far (e.g. use case diagrams [8], or sequence diagrams, cf. [29]). There is only little work on the Business Process Model and Notation (see [5]), and even less on UML activity diagrams [20].

A detailed discussion of aesthetic criteria for class diagrams is found in [7, p. 54–65], a recent survey of empirical results on layout criteria is found in [9]. Wong and Sun [29] provide an overview of these criteria from a cognitive psychology point of view, along with an evaluation of how well these principles are realized in several UML CASE tools. Purchase et al. discuss aesthetic criteria with a view to the layout of UML class and communication diagrams (cf. [18, 17]) and also provide sources to justify and explain these criteria (cf. [19]). Eichelberger [6] also discusses these criteria at length, and shows how they can be used in the automatic layout of UML class diagrams.

In order to develop automatic layout algorithms that are perceived as good by human modelers, detailed knowledge about the individual criteria, their relative and absolute impact, and their formalization is needed. So, it is not surprising that most of the empirical research on UML diagrams has so far focused on studying individual principles, with an emphasis on the second group (cf. [21, 7, 10, 30, 18]). For instance, work by Purchase et al. has shown that there are many such criteria with varying degrees of impact (see e.g. [18]), though all of them seem to have a rather small impact with findings that are not highly or not at all statistically significant. Also, the ranking and contribution of these criteria may vary across different diagram types. Even between class and communication diagrams, which are rather close relatives as far as concrete syntax is concerned, [18, pp. 246] shows notable differences in the ordering and impact of layout criteria. Thus, other notations that share even less commonalities with class diagrams (e.g., activity, use case, or sequence diagrams) may need a completely different set of criteria.

7. CONCLUSIONS

Previously, we found that layout quality does impact the understanding of UML diagrams [24], irrespective of diagram type but dependent on modeler expertise [25]. We also found that diagram size had a significant influence [26], but we could so far not tie our findings to diagram *quality* because (a) there was no such metric, and (b) our size metric encompassed some aspects of quality, resulting in three questions.

- **RQ 1:** Do the results of our previous study hold up when analyzing a (much) larger, more diverse data set?
- **RQ 2:** Diagram quality is an elusive notion: can we characterize it in a precise way, at least part of it?
- **RQ 3:** What is the impact of diagram quality on model understanding, as compared to diagram size?

Regarding **RQ1**, we refined our existing notion of diagram size, and removed quality aspects. Three independent assessors applied the metric to 60 diagrams and yielded results with very high correlation. Thus, the metric definition is now sufficiently precise. We repeated our previous analysis, and despite minor variations, earlier results were confirmed.

Regarding **RQ2**, we developed a metric for topographic layout quality. It includes all known quality aspects backed by empirical data. We validated the metric by comparing the results of three independent raters (correlation 0.97).

Regarding **RQ3**, we have correlated diagram size and quality with diagram types and expertise levels. We found that diagram size has the expected effect (a negative correlation) on model understanding. However, it is a little smaller than reported previously [26]. We also find that diagram quality (as defined here) has the same effect, but much more so. This is very intuitive given that the previous definition of size *included* some aspect of quality. We derived (rough) recommendations for the size and quality of diagrams in terms of the number of elements, flaws, and their ratio.

The validity of our findings depends on three factors. Firstly, it depends on the reliability of the underlying experimental data. To our best knowledge, this data set is the largest of its kind, and great care has been taken to ensure the methodological soundness of the underlying experiments.

Secondly, our findings depend on whether the metrics we have proposed do indeed capture size and quality of diagrams adequately. We collected diagram quality aspects from the literature, and there is little doubt that they all are relevant. Some, however, are not very widely studied, in particular qualities relating to flow and symmetry. While these aspects are certainly important, they are difficult to formalize, and there is currently not much empirical data available about them. So, this aspect has to be deferred to future work.

Thirdly, it is crucial whether the diagram sample used can be considered representative. Given that there is no similar body of data available, it is difficult to establish this as a fact. One thing that is known, is that the five diagram types used in our study represent the most used UML diagram types, so our study is representative at least in this respect. However, the number of diagrams per diagram type (twelve) is relatively small (although larger than in any other published study). To address this concern, we have analyzed our sample with regards to the distributions of size and quality. Where reference data is available, we have compared them and found our sample representative.

While this study is certainly not the last word on the issue of diagram layout quality, we believe it offers more validity than comparable studies on UML diagrams. Nevertheless, more research is needed to refine and independently replicate our findings. In order to facilitate that, we have published all our experimental material online together with this paper, along with the raw data at http://bit.ly/1RJrv8K.

8. REFERENCES

[1] H. Allder and D. Carrington. Graph Layout Aesthetics in UML Diagrams: User Preferences. *Graph algorithms and applications 3*, page 255, 2004.

[2] G. Di Battista, P. Eades, R. Tamassia, and I. G. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs.* Prentice Hall, 1999.

[3] Brian Dobing and Jeffrey Parsons. How UML is used. *Comm. ACM*, 49(5):109–113, 2006.

[4] T. Dwyer, B. Lee, D. Fisher, K. Quinn, P. Isenberg, G. Robertson, and C. North. A Comparison of User-Generated and Automatic Graph Layouts. *IEEE Txn. Visualization and Computer Graphics*, 15(6):961–968, 2009.

[5] P. Effinger, N. Jogsch, and S. Seiz. On a Study of Layout Aesthetics for Business Process Models Using BPMN. In *Proc. 2nd Intl. Ws. Business Process Modeling Notation*, pages 31–45. Springer, 2010.

[6] H. Eichelberger. Aesthetics of class diagrams. In *Proc. 1st Intl. Ws. Visualizing Software for Understanding and Analysis (VISSOFT)*, pages 23–31. IEEE, 2002.

[7] H. Eichelberger. *Aesthetics and automatic layout of UML class diagrams.* PhD thesis, University of Würzburg, 2005.

[8] H. Eichelberger. Automatic layout of UML use case diagrams. In *Proc. 4th ACM Symp. Sw. Visualization (SOFTVIS)*, pages 105–114. ACM, 2008.

[9] H. Eichelberger and K. Schmid. Guidelines on the aesthetic quality of UML class diagrams. *Information and Software Technology*, 51(12):1686–1698, 2009.

[10] M. Eiglsperger. *Automatic layout of UML class diagrams: a topology-shape-metrics approach.* PhD thesis, Univ. Tübingen, 2003.

[11] H. Eysenck. Speed of Information Processing, Reaction Time, and the Theory of Intelligence. In P. Vernon, editor, *Speed of Information-Processing and Intelligence*, chapter 2, pages 21–67. Ablex Publishing Corp., 1987.

[12] C. Gurr. Effective Diagrammatic Communication: Syntactic, Semantic and Pragmatic Issues. *J. Visual Languages and Computing*, 10:317–342, 1999.

[13] B. Karasneh and M. Chaudron. Online Img2UML Repository: An Online Repository for UML Models. In *Intl. Ws. Experiences and Empirical Studies in Software Modelling (EESSMod)*, 2013. co-located MoDELS'13, repository available at http://cse-poros.cse.chalmers.se.

[14] P. Langer, T. Mayerhofer, M. Wimmer, and G. Kappel. On the Usage of UML: Initial Results of Analyzing Open UML Models. In H.-G. Fill, D. Karagiannis, and U. Reimer, editors, *Proc. Modellierung*, pages 289–304. Gesellschaft für Informatik, 2014.

[15] J. Oberlander. Grice for Graphics: Pragmatic Implicature in Network Diagrams. *Information Design Journal*, 8(2):163–179, 1996.

[16] H. Purchase. Metrics for Graph Drawing Aesthtetics. *J. Visual Languages and Computing*, 13(5):501–516, 2002.

[17] H. Purchase, J. Allder, and D. Carrington. Graph layout aesthetics in UML diagrams: user preferences. *J. Graph Algorithms Applications*, 6(3):255–279, 2002.

[18] H. Purchase, D. Carrington, and J. Allder. Empirical Evaluation of Aesthetics-based Graph Layout. *J. Empirical Software Engineering*, 7(3):233–255, 2002.

[19] H. Purchase, L. Colpoys, D.A. Carrington, and M. McGill. UML Class Diagrams: An Emprical Study of Comprehension. In Kang Zhang, editor, *Software-Visualization: From Theory to Practice*, pages 149–178. Kluwer, 2003.

[20] G. Reggio, F. Ricca, G. Scanniello, F. Di Cerbo, and G. Dodero. On the comprehension of workflows modeled with a precise style: results from a family of controlled experiments. *Software & Systems Modeling*, pages 1–24, 2013.

[21] J. Seemann. Extending the Sugiyama algorithm for drawing UML class diagrams: Towards automatic layout of object-oriented software diagrams. In *Proc. Intl. Conf. Graph Drawing (GD)*, pages 415–424. Springer, 1997.

[22] B. Sharif and J. Maletic. An eye tracking study on the effects of layout in understanding the role of design patterns. In *Proc. 2010 IEEE Intl. Conf. Software Maintenance (ICSM)*, pages 41–48. IEEE, 2010.

[23] B. Sharif and J. Maletic. The Effects of Layout on Detecting the Role of Design Patterns. In *Proc. 23rd IEEE Conf. Software Engineering Education and Training (CSEE&T)*, pages 41–48. IEEE, 2010.

[24] Harald Störrle. On the Impact of Layout Quality to Unterstanding UML Diagrams. In *Proc. IEEE Symp. Visual Lang. and Human-Centric Computing (VL/HCC)*, pages 135–142. IEEE CS, 2011.

[25] Harald Störrle. On the Impact of Layout Quality to Unterstanding UML Diagrams: Diagram Type and Expertise. In Gennaro Costagliola and others, editors, *Proc. IEEE Symp. Visual Languages and Human-Centric Computing (VL/HCC)*, pages 195–202. IEEE CS, 2012.

[26] Harald Störrle. On the Impact of Layout Quality to Understanding UML Diagrams: Size Matters. In Jürgen Dingel and others, editors, *Proc. 17th Intl. Conf. Model Driven Engineering Languages and Systems (MoDELS)*, number 8767 in LNCS, pages 518–534. Springer Verlag, 2014.

[27] J. Swan, M. Kutar, T. Barker, and C. Britton. User Preference and Performance with UML Interaction Diagrams. In *Proc. 2004 IEEE Symp. Visual Languages and Human Centric Computing (VL/HCC)*, pages 243–250. IEEE, 2004.

[28] C. Ware. *Information Visualization: Perception for Design.* Morgan Kaufmann, 2nd edition, 2004.

[29] K. Wong and D. Sun. On evaluating the layout of UML diagrams for program comprehension. *Software Quality J.*, 14(3):233–259, 2006.

[30] S. Yusuf, H. Kagdi, and J. Maletic. Assessing the Comprehension of UML Class Diagrams via Eye Tracking. In *15th IEEE Intl. Conf. Program Comprehension*, pages 113–122. IEEE CS, 2007.