Review Time as Predictor for the Quality of Model Inspections

Marian Daun[1][a], Meenakshi Manjunath[1][b] and Jennifer Brings[2][c]
[1]*Center of Robotics, Technical University of Applied Sciences Würzburg-Schweinfurt, Schweinfurt, Germany*
[2]*University of Duisburg-Essen, Essen, Germany*

Keywords: Validation, Software Inspection, Model Inspection, Ad Hoc Review, Review Time, Controlled Experiment.

Abstract: Software inspections play an important part in ensuring the quality of software development. With the emergence of model-based development approaches, there is also a need for model inspections to ensure correctness of model-based artifacts. In practice, ad hoc inspections are regularly conducted, often by new and rather inexperienced colleagues, which are asked spontaneously to review an artifact of interest. The use of novices, such as trainees or student assistants, allows shorter review cycles at reduced costs. The quality of these ad hoc inspections is commonly attributed to different factors, often related to the reviewer. Increasing review time can be seen as an indicator that the reviewer takes the review serious. Furthermore, with more time spent, it can be assumed that more defects will be found. In this paper, we report the results of an experiment on ad hoc model inspections. Our results show that – contradictory to these assumptions and empirical findings from inspections of textual documents – the review time a reviewer decides to spend on a review has no significant influence on the effectiveness of ad hoc model inspections.

1 INTRODUCTION

During software engineering processes, manual quality assurance is regularly mandated and conducted at different stages and with different intensities (ISO 26262-1, 2011; ISO/IEC 25030, 2007). While formal inspections are completed from time to time, on multiple occasions brief visual inspections by coworkers or the developers themselves are done regularly. These visual reviews of the requirements (Miller et al., 1998), the code (de Almeida et al., 2003), or other development artifacts (Laitenberger et al., 2000) aim at improving the overall quality of the software product to be developed.

In the past, research has been conducted on software inspections and other formal validation techniques. On a regular basis, different validation techniques have been compared to ad hoc reviews, often showing that they are more effective and efficient. However, most validation techniques exceed ad hoc reviews in the resources needed (e.g., since multiple reviewers are expected to be present, multiple intense reviewing days are defined), which leads to increased costs. Hence, such validation techniques are used at distinct points but not commonly throughout a development project. Therefore, there is still a need for ad hoc reviews.

Ad hoc reviews are often conducted with inexperienced reviewers, these are typically available student assistants working in the company unit or newer colleagues or trainees that have not yet assigned to specific duties. As they tend to be more error-prone than the use of - pricey and seldomly available - experts in the field, often ad hoc reviews are distributed among a larger set of reviewers. However, there is a need to assess the quality of the inspection results, as the developers typically want to focus on the more reliable reviews that provide less false positives. Established research identify three major factors influencing the quality of an inspection:

1. The reviewer, i.e. personal factors such as years of experience, degrees achieved, etc.
2. The review subject, i.e. the way the requirements, code, models, other artifacts under review have been prepared and are structured.
3. The review process, i.e. the technique applied, the number of resources and time spent.

A common misconception about reviews is that the time spent for reviewing (i.e. the *review time*) influences the quality of a review. Particularly, it

[a] https://orcid.org/0000-0002-9156-9731
[b] https://orcid.org/0009-0005-6421-1450
[c] https://orcid.org/0000-0002-2918-5008

Daun, M., Manjunath, M. and Brings, J.
Review Time as Predictor for the Quality of Model Inspections.
DOI: 10.5220/0012682900003687
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 19th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2024)*, pages 120-131
ISBN: 978-989-758-696-5; ISSN: 2184-4895
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

is assumed that ad hoc reviews suffer from too little time spent whereas on-the-fly reviewing a co-worker's code, text, model, etc. This means that reviews conducted in less time are usually less effective and efficient when compared to reviews conducted in more time. If that is the case, then, we could prioritize reviews with more review time taken higher than reviews with less time taken.

In this paper, we investigate this effect for visual inspections of models (i.e. *ad hoc model inspections*), in contribution to the research goal: *Does review time influence ad hoc model inspections?*

At this point, we briefly lay out our definition of review time. Review time refers to the duration taken by a reviewer to conduct a review task. However, we do not refer to predefined review times, i.e. a reviewer is asked to spend two hours for a review. Instead, we refer to the actual time used without given restrictions. This implies that we want to find out the influences of the actual time a reviewer deems sufficient for a reviewing task or a reviewer is willing to invest for a review on the quality of the review. This approach makes it challenging to offer straightforward guidelines on the recommended duration for an uncompromised quality review. In contrast, we want to contribute to the question, whether the time taken is an indicator for finding good reviewers or to estimate whether a review is more likely to be of good quality or not.

To answer the research goal, an experiment on the influence of review time on ad hoc model inspections was conducted and is reported in this paper. In total, 200 participants conducted ad hoc model inspections. In total, we collected data from 520 participants who performed multiple ad hoc model inspection tasks across a total of eight different tasks. After filtering, 497 data sets were used for analyzing the influence of review time on effectiveness, reviewers' confidence, and efficiency of ad hoc model inspections. The results show that there is no discernible effect for effectiveness, while confidence and efficiency are influenced by review time. However, results show that increased review time does neither lead to increased efficiency and confidence, nor does decreasing review time. Moderate review time leads to significantly higher efficiency and confidence compared to very short and very long review time.

This paper is structured as follows. Section 2 introduces background information and related work on ad hoc model inspections and related studies. Subsequently, Section 3 introduces the study design and Section 4 the study results. The major findings and threats to validity are discussed in Section 5. Finally, Section 6 concludes the paper.

2 RELATED WORK

2.1 Ad Hoc Model Inspections

Different inspection techniques have been proposed to support validation of various software development artifacts. Among others, formal inspection ((Fagan, 1976; Fagan, 1986), often referred to as Fagan-Inspection), walkthroughs (Boehm, 1987), N-fold inspection (Martin and Tsai, 1990), checklist-based inspection (Thelin et al., 2003), perspective-based reading (Shull et al., 2000) and scenario-based reading (Regnell et al., 2000) have gained much attention and are regularly investigated for their effectiveness and efficiency (e.g., (Miller et al., 1998; Basili et al., 1996)). This is commonly done by comparing these techniques between each other, or even frequently by comparison with ad hoc inspections (or ad hoc reviews). Ad hoc inspections are typically defined as inspections that are conducted without any guidance for the reviewer and without a prescribed process. Basically, the reviewer is just given the review artifact and the task to validate its correctness (Porter et al., 1995; O.Oladele and O. Adedayo, 2014).

The majority of existing studies is interested in inspecting requirements artifacts or code artifacts. The inspection of requirement artifacts is, for instance, investigated by Miller et al. (Miller et al., 1998). In a controlled experiment, trained student participants conduct inspections for error detection in natural language requirements specifications. Basili et al. 1996 (Basili et al., 1996) report on a controlled experiment with professional software developers comparing different inspection techniques for requirements documents. Finding out that perspective-based review is significantly more effective than other inspection techniques for requirements documents. Other examples for requirements inspection studies were conducted by He and Carver (He and Carver, 2006), Maldonado et al. (Maldonado et al., 2006), Laitenberger et al. (Laitenberger et al., 2001), Berling and Runeson (Berling and Runeson, 2003), and Sabaliauskaite et al. (Sabaliauskaite et al., 2004), which often come to comparable findings. Code inspections are, among others, studied by Porter et al. (Porter et al., 1997), Laitenberger (Laitenberger, 1998), Almeida et al. (de Almeida et al., 2003), or Dunsmore et al. (Dunsmore et al., 2003).

However, while requirements inspections and code inspections have been heavily investigated, model inspections are also the center of various studies. For instance, de Mello et al. (d. Mello et al., 2012) investigate the inspection of feature models. Conradi et al. (Conradi et al., 2003) and Laitenberger et al. (Laitenberger et al., 2000) report experiments on inspections of UML models.

In previous work, we proposed dedicated review models to improve model inspections (Daun et al., 2014). Results showed that Message Sequence Charts are a favorable modeling language for conducting reviews. Particularly, we conducted experiments to compare the use of Message Sequence Charts with functional specification languages (Daun et al., 2019b), of review models merging multiple specifications (Daun et al., 2019a), of the representation format for inconsistencies shown in the review model (Daun et al., 2017), or the use of instance- vs. type-level specifications (Daun et al., 2020). By analyzing the data gathered from all these experiments, we tried to identify predictors for a reviewer's performance, but concluded so far that commonly suggested predictors like experience and confidence are not reliable for the quality of model inspections (Daun et al., 2021).

In summary, ad hoc inspections are regularly used as comparison for more advanced inspection techniques under investigation. Although other inspection techniques typically win the comparison with ad hoc inspections, controlled experiments do exist that found out that ad hoc inspections do not perform worse than systematic inspection techniques (Lanubile and Visaggio, 2000) or at least not worse than all other systematic inspection techniques (Porter et al., 1995; Porter and Votta, 1998). In addition, ad hoc model inspections possess benefits regarding the low resource consumption. This allows conducting ad hoc inspections frequently whenever validation is needed. Therefore, ad hoc inspections are regularly used in industry (although typically not as the only inspection technique used during the entire project).

2.2 Influence Factors

Beside experiments comparing effectiveness and efficiency of different inspection techniques, studies exist aiming at investigating other influence factors for inspections. For instance, a variety of studies investigates the influence the used notation has. Particular emphasis is typically given to the set of symbols used. Figl et al. (Figl et al., 2013a) investigate the influence of the symbol sets used in modeling languages. In a study with 136 participants, it is shown that perceptual distinctiveness and semiotic clarity of the used symbols affects model comprehension. Particularly, the correctness of model understanding, the cognitive load to be possessed, and the time needed for understanding the models varies significantly. In (Figl et al., 2013b), Figl et al. report another study with 155 student participants, showing that aesthetic design of the used notational elements can improve the model understanding of process models. Nugroho conducted an experiment with graduate students, finding out that for UML diagrams, the level of detail has a significant influence (Nugroho, 2009). Lucia et al. report in (Lucia et al., 2008) results of two controlled experiments with Bachelor and Master students showing that UML class diagrams are significantly easier to comprehend than ER diagrams. Bavota et al. conducted a study to compare UML class diagrams and ER diagrams regarding their impact on model comprehension (Bavota et al., 2011). They also showed that UML class diagrams are in general easier to comprehend than ER diagrams.

Further studies, are more broadly looking at other influence factors than modeling language related issues. Mendling et al. report in (Mendling et al., 2012) a study investigating the influence of model and personal factors on the comprehension of process models. Major findings are that comprehension is hindered by the annotation of additional semantic information, and that theoretical knowledge as well as modeling experience support model comprehension. In (Zimoch et al., 2017), Zimoch et al. report a study investigating the influence of process modeling experience on model comprehension. They conclude that experience in general has a positive impact on model comprehension. However, in case complexity of the models under investigation is considerably increased, the impact of experience vanishes more and more.

In conclusion, literature attributes three major factors influencing effectiveness and efficiency of model inspections: (a) the inspection technique applied, (b) syntax related issues of the inspected model, and (c) experience as personal factor.

Although time has not yet been widely investigated, existing studies on these three factors, partly, also report an influence of time needed. As this was in no case the major point of investigation, findings are typically only briefly summarized. From these findings, it can be concluded that time influences the model inspection in such a way that the more time needed, the better the inspection result. This, particularly, often interwoven with the investigation of the inspection technique (e.g., (Lanubile and Visaggio, 2000)). This means that in many cases it has been found out that inspection techniques taking more time are advantageous. For instance, perspective-based reading is more time-consuming than checklist-based inspections and also found to be more effective (e.g., (Basili et al., 1996)). However, review time is perceived as something costly, that should be minimized in industrial practice (cf. e.g., (Doolan, 1992)).

3 STUDY DESIGN

For experiment reporting, we keep to established best practices (Wohlin et al., 2000; Jedlitschka et al., 2008), which helps, among others, increase comprehensibility and comparability with other experiments.

3.1 Goal and Research Questions

As stated in Section 1 the overall goal of this study is to investigate whether review time does influence ad hoc model inspections. To achieve this goal, we investigate the effects review time has on effectiveness, confidence, and efficiency of ad hoc model inspections. Therefore, we define three research questions:

RQ1: Does review time influence the effectiveness of ad hoc model inspections?
RQ2: Does review time influence the reviewers' confidence in ad hoc model inspections?
RQ3: Does review time influence the efficiency of ad hoc model inspections?

3.2 Variables

Review Time: is measured in seconds and is defined as the time used for ad hoc model inspection of one model. As review time is defined on an open-end ratio scale, we also define review time intervals to allow for better comparison of means. Therefore, the ratio scale is transferred into an ordinal scale.[1]
Effectiveness: is measured as the ratio of correct review decisions made compared to all review decisions made. Hence, effectiveness is measured on a ratio scale from 0 (i.e. 0% correct decisions made) to 1 (i.e. 100% correct decisions made).
Confidence: is defined as the average confidence the reviewer claims for the review decisions made. Confidence is measured on 5-point semantic differential scale, where 1 means very unconfident and 5 very confident. However, as confidence is calculated as mean of all review decisions made for a model, confidence is defined on a ratio scale from 1 to 5.
Efficiency: is measured as the average time used for a correct decision made. Efficiency is measured in seconds and defined on an open-end ratio scale. Note that efficiency is not independent of review time. Nevertheless, we are interested in efficiency, as industry typically aims at efficient reviews (cf. (Doolan, 1992)). Thus, it is of interest to determine whether a high efficiency is bound to certain ranges of review time.

3.3 Hypotheses

Based on the research questions, we define the following null and alternative hypotheses:

$H1_0$: *There is no effect of review time on effectiveness.*
$H1_{A1}$: *Increasing review time leads to increased effectiveness.*
$H1_{A2}$: *Increasing review time leads to decreased effectiveness.*
$H1_{A3}$: *Review time influences effectiveness, but the effect is not linear.*

$H2_0$: *There is no effect of review time on confidence.*
$H2_{A1}$: *Increasing review time leads to increased confidence.*
$H2_{A2}$: *Increasing review time leads to decreased confidence.*
$H2_{A3}$: *Review time influences confidence, but the effect is not linear.*

$H3_0$: *There is no effect of review time on efficiency.*
$H3_{A1}$: *Increasing review time leads to increased efficiency.*
$H3_{A2}$: *Increasing review time leads to decreased efficiency.*
$H3_{A3}$: *Review time influences efficiency, but the effect is not linear.*

3.4 Participants

The experiment was conducted with student participants. The students are mostly graduate students enrolled in degree programs for applied computer science and business information systems. Participants were recruited in software engineering courses. Due to the courses' syllabi, it was ensured that the participants do have sufficient knowledge of validation activities and the modeling languages investigated, in addition, they were trained to conduct ad hoc model inspections for these kinds of models. In total, 200 students participated in the experiment. As each participant conducted multiple ad hoc model inspections, a total of 520 data sets were collected.

[1] Note that while we typically use five-minute intervals, the scale is no interval scale, as we use a catch-all group for all review times greater than thirty-five minutes.

3.5 Experiment Material

As experiment material, excerpts from industrial specifications have been used. These have been revised to match intended size and complexity, to remove intellectual property as well as issues relating to needed in-depth domain expertise. Models were chosen to fit approximately one page. As modeling languages, Message Sequence Charts (International Telecommunication Union, 2016), automata (de Alfaro and Henzinger, 2001), and functional architecture models (Albers et al., 2016) have been used.

3.6 Experiment Design and Procedure

The experiment was conducted online and was designed to last about 30–40 minutes, in which the participants conducted multiple ad hoc model inspections back to back. In addition, a post hoc questionnaire was used to collect demographic data.

3.7 Analysis Procedure

The data sets were filtered for data sets indicating non-serious participation (i.e. 23 data sets were removed). The remaining data sets were analyzed by calculating common descriptive statistic parameters.

To estimate overall influence of review time on effectiveness, confidence, and efficiency, Pearson's correlation and simple regression were conducted. Therefore, the original review time was used.

As mentioned above, we sorted review time into intervals. Review time intervals (1-5 minutes, 5–10 minutes, 10–15 minutes, 15–20 minutes, 20–25 minutes, 25–30 minutes, 30–35 minutes, and >30 minutes). These review time intervals were used for conducting one-way independent analysis of variance (ANOVA) to compare effectiveness, confidence, and efficiency for different review time. In addition, Levene's test was used to determine difference of variances. In case of heterogeneous variances, the Welsh-Test was conducted. Post hoc analyses for significant ANOVA results included the Bonferroni-Test and the Games-Howell-Test.

4 RESULTS

4.1 Descriptive Statistics

The descriptive statistics for explanatory and response variables are given in Table 1. In addition, Fig. 1 shows the distribution of the data. We excluded 23 data sets from the investigation. In these 23 cases, the ad hoc model inspection was conducted in less than one minute. Investigation of the results substantiated the assumption that this means that participants did not partake seriously. As can be seen for effectiveness and confidence, values are distributed across the entire defined scale. In mean ad hoc model inspection resulted in 56% correct review decisions made and a confidence of 3.26, which is above an expectation value of 3. In mean, participants needed 591 second (i.e. almost 10 minutes) for an entire review and 48 seconds for a correct decision made.

Table 1: Descriptive Statistics.

		Effectiveness	**Confidence**	**Efficiency**	**Review Time**
N	Valid	497	496	497	497
	Miss.	23	24	23	23
Mean		0.561	3.260	47.605	591.21
Std. Err.		0.009	0.046	5.257	25.264
Median		0.571	3.417	4.511	463
Std. Dev.		0.21	1.029	117.202	563.218
Variance		0.044	1.059	13736.368	317214.028
Min.		0	1	0	61
Max.		1	5	1310	5973
Percentiles	25	0.417	2.333	1.849	257.5
	50	0.571	3.416	4.511	463
	75	0.714	4	51.3	756

Table 2, shows descriptive statistics for the effectiveness, confidence, and efficiency in relation to the defined review time intervals. As can be seen, most ad hoc model inspections took 5–10 minutes, while the absolute majority of reviews were conducted within 1–15 minutes. The distribution for the review time intervals is visualized using box plots in Fig. 2.

4.2 Hypotheses Tests

4.2.1 Effectiveness

For analyzing the influence of review time on the effectiveness of ad hoc model inspections, we conducted Pearson correlation, simple regression, and analysis of variance.

Pearson correlation shows no correlation between effectiveness and review time ($r = .033, p = .461$). A simple regression shows no significant regression equation ($F(1,495) = .545, p = .461$), with $R^2 = -.001$. Thus, review time cannot be used to explain effectiveness.

To investigate the effect of review time on effectiveness, we also conducted a one-way independent ANOVA. As outlined in Section 4.1, we grouped review time in intervals of five minutes to investigate intergroup effects. There was no significant effect of review time on effectiveness ($F(7,489) = .973, p = .450$). Levene's test indicated equal variances ($F = .997, p = .437$), thus we assume the ANOVA reliable. Consequently, we cannot reject $H1_0$.

Table 2: Descriptive Statistics for Effectiveness, Confidence, and Efficiency Depending on Review Time in Minutes.

						95% Conf. Interval			
	Review Time	N	Mean	Std. Dev.	Std. Err.	Lower	Upper	Min.	Max.
Effectiveness	*1-5*	151	0.576	0.214	0.017	0.542	0.610	0.000	0.917
	5-10	159	0.553	0.213	0.017	0.519	0.586	0.083	1.000
	10-15	108	0.562	0.211	0.020	0.522	0.603	0.083	1.000
	15-20	44	0.518	0.173	0.026	0.465	0.570	0.000	0.917
	20-25	13	0.521	0.234	0.065	0.380	0.662	0.167	0.857
	25-30	9	0.598	0.217	0.072	0.431	0.765	0.333	0.917
	30-35	6	0.706	0.230	0.094	0.465	0.948	0.333	1.000
	>35	7	0.605	0.164	0.062	0.453	0.756	0.375	0.750
	Total	497	0.561	0.210	0.009	0.543	0.580	0.000	1.000
Confidence	*1-5*	150	2.693	0.960	0.078	2.539	2.848	1.000	5.000
	5-10	159	3.271	1.047	0.083	3.107	3.435	1.167	5.000
	10-15	108	3.749	0.806	0.078	3.595	3.902	1.429	5.000
	15-20	44	3.775	0.583	0.088	3.598	3.952	2.583	5.000
	20-25	13	3.449	0.841	0.233	2.941	3.957	1.917	5.000
	25-30	9	4.163	0.667	0.222	3.650	4.675	3.167	5.000
	30-35	6	4.169	0.579	0.236	3.561	4.777	3.143	4.750
	>35	7	2.104	1.165	0.440	1.027	3.181	1.250	3.833
	Total	496	3.260	1.029	0.046	3.169	3.351	1.000	5.000
Efficiency	1-5	151	34.687	24.744	2.014	30.708	38.665	0.000	144.500
	5-10	159	44.364	60.634	4.809	34.867	53.862	0.539	359.000
	10-15	108	27.425	65.453	6.298	14.939	39.910	0.932	366.000
	15-20	44	29.345	77.977	11.755	5.638	53.052	0.000	357.667
	20-25	13	20.147	55.942	15.516	-13.659	53.953	2.017	206.167
	25-30	9	41.493	111.426	37.142	-44.157	127.142	2.308	338.600
	30-35	6	167.394	181.572	74.127	-23.155	357.942	3.783	392.000
	>35	7	782.188	439.677	166.182	375.554	1188.821	9.014	1310.000
	Total	497	47.605	117.202	5.257	37.276	57.934	0.000	1310.000

4.2.2 Confidence

As for effectiveness, we conducted Pearson correlation, simple regression, and analysis of variance.

Confidence is positively related to review time. Pearson correlation shows a small effect of $r = .169$ that is highly significant at $p < .001$. A simple regression was calculated to predict confidence based on review time. A significant regression equation was found ($F(1,494) = 14.486, p < .001$), with a small $R^2 = .028$.

A one-way independent ANOVA shows a significant effect of review time on confidence ($F(7,488) = 18.039, p < .001$). As Levene's test indicated unequal variances ($F = 6.007, p < .001$), we conducted the Welch-Test, which confirmed the findings from the ANOVA ($F(7,35.834) = 19.716, p < .001$). Hence, we can reject $H2_0$ and accept $H2_A$. We used post hoc tests to investigate the differences between the groups, while different tests yielded in comparable results, we focus on the results of the Games-Howell-Test as it meets the preconditions best (see Figure 3).

First, the two groups with the least time consumption significantly differ from groups with moderately more time consumption: Confidence for a review time of 1–5 minutes ($M = 2.693, SD = 0.96$) significantly differs from a review time of 5–10 minutes ($M = 3.271, SD = 1.047$), 10–15 minutes ($M = 3.749, SD = 0.806$), 15–20 minutes ($M = 3.775, SD = 0.583$), 25–30 minutes ($M = 4.163, SD = 0.667$), and 30–35 minutes ($M = 4.169, SD = 0.579$). In addition, confidence for a review time of 5–10 minutes ($M = 3.271, SD = 1.047$) also significantly differs from a review time of 10–15 minutes ($M = 3.749, SD = 0.806$), 15–20 minutes ($M = 3.775, SD = 0.583$), and 25–30 minutes ($M = 4.163, SD = 0.667$).

Second, Confidence for a review time of more than 35 minutes ($M = 2.104, SD = 1.165$) significantly differs from a review time of 25–30 minutes ($M = 4.163, SD = 0.667$) and 30–35 minutes ($M = 4.169, SD = 0.579$). Hence, confidence is not increasing with increasing review time. While this is the case for shorter review time, for longer review time confidence is decreasing. Therefore, we can neither accept $H2_{A1}$, nor $H2_{A2}$, but accept $H2_{A3}$.

4.2.3 Efficiency

Again, we conducted Pearson correlation, simple regression, and analysis of variance.

Efficiency is positively related to review time by a large effect of $r = .602$. The effect is highly significant at $p < .001$. A simple regression found a significant regression equation ($F(1,495) = 280.887, p < .001$) with $R^2 = .362$. Participants' predicted efficiency is equal to $-26.418 + .125(\mathit{ReviewTime})$ seconds/correct answer when review time is measured in seconds.

A one-way independent ANOVA shows a significant effect of review time on efficiency ($F(7,489) = 96.898, p < .001$). However, Levene's test indicated unequal variances ($F = 38.88, p < .001$). Therefore, we conducted the Welch-Test, which confirmed the findings from the ANOVA ($F(7,34.57) = 3.819, p =$

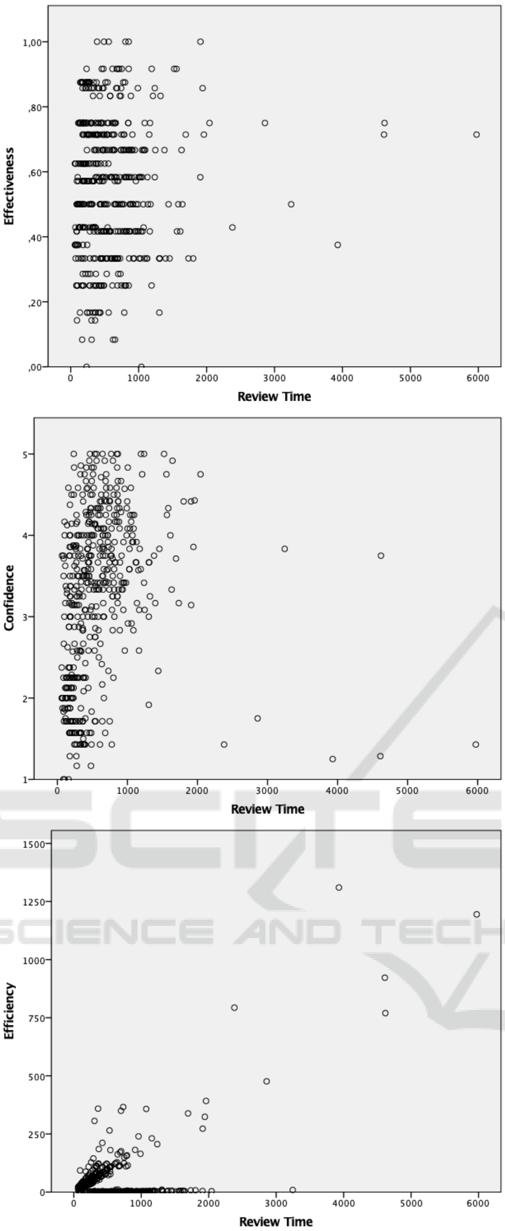

Figure 1: Scatterplots.

.004). Thus, we can reject $H3_0$ and accept $H3_A$. We used post hoc tests to investigate the differences between the groups, while different tests yielded in comparable results, we only report results of the Games-Howell-Test. Significant differences in efficiency do only exist between large review time (i.e. ¿35 min) and lower review time. Namely: Efficiency for a review time of more than 35 minutes ($M = 782s, SD = 440s$) significantly differs from a review time of 1–5 minutes ($M = 35s, SD = 25s$), 5–10 minutes ($M = 44s, SD = 61s$), 10–15 minutes ($M = 27s, SD = 65s$), 15–20 minutes ($M = 29s, SD = 78s$), 20–25 minutes ($M = 20s, SD = 56s$), and 25–30 minutes ($M = 41s, SD = 111s$).[2]

Thus, efficiency is not increasing with increasing review time. While this is the case for lower review time, for large review time efficiency is decreasing. Therefore, we reject $H3_{A1}$. Considering the significant results, we accept $H3_{A2}$. Taking the increasing means for lower review time into account, we also accept $H3_{A3}$.

5 DISCUSSION

5.1 Major Findings

With respect to the three investigated research questions, we can conclude three major findings regarding the influence of review time on ad hoc model inspections:

Regarding **RQ1**, we found out that review time does not influence the effectiveness of ad hoc model inspections. This is based on the absence of significant correlations or regression, and that one-way independent ANOVA did not yield significant results.

Regarding **RQ2**, we can state that review time does have an influence on the reviewers' confidence of ad hoc model inspections. Very small effects are found using correlation and regression analyses. In addition, analysis of variance showed a significant difference between groups. We found out that while in principle confidence is increasing with increasing review time, a very long review time results in the lowest confidence. Regarding significance, it can be stated that review times of about 10–35 minutes for one model to be investigated during ad hoc model inspection related to a higher confidence than review times of 1–10 minutes and above 35 minutes.

Regarding **RQ3**, we found out that review time does influence efficiency of ad hoc model inspections. Correlation and regression analysis found large statistically significant effects, which is not surprising considering the inherent relationship between review time and efficiency. However, analysis of variance shows a more fine-grained view. Large review times of more than 35 minutes (considering the Bonferroni-Test, maybe also of 30–35 minutes) result in significantly less efficient ad hoc model inspections.

[2] Note that other tests, such as the Bonferroni-Test, find also significant differences between more than 35 minutes and 30–35 minutes, and between 30–35 minutes and all other groups. However, due to the heterogeneity of variances, we keep to the interpretation of Games-Howell-Test, although differences of means are indeed large for these group comparisons.

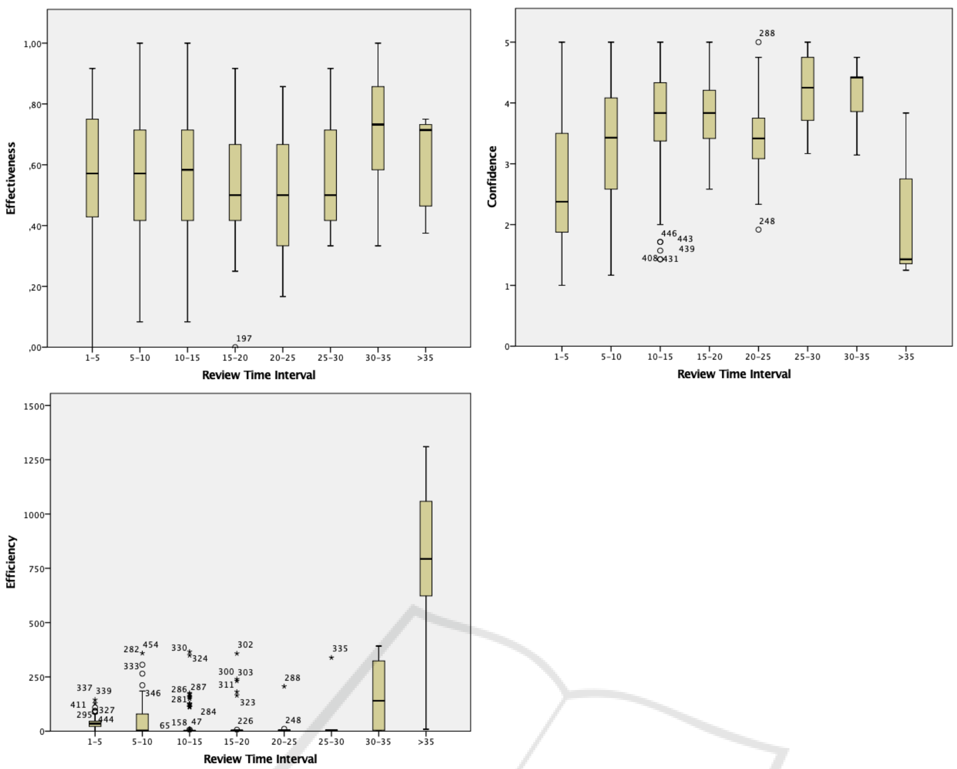

Figure 2: Box Plots.

In summary, we can confirm that review time has an influence on ad hoc model inspections. However, contrary to other investigations on influencing factors on the inspections, we could not find any influence of review time on the effectiveness of the review. Regarding the influence on the reviewer's confidence and the efficiency of the review, we also found evidence that contradicts assumptions from the related work. Increasing review time does not necessarily lead to better reviews (i.e. reviews, where the reviewer is more confident in decision-making and that are conducted with a higher efficiency). There seems to be a point where more review time leads to worse reviews (in terms of confidence and efficiency). Nevertheless, for small and moderate review times of up to 30 minutes for inspecting one model, we can substantiate claims that increasing review time leads to better reviews (although only in terms of confidence and efficiency but not for effectiveness).

5.2 Threats to Validity

5.2.1 Threats to Internal Validity

In online experiments, a threat of participants losing interest, dropping out of the experiment, or corrupting the measurements by idling must always be considered. To lower this threat, we designed the experiment such that two to three ad hoc model inspections could be conducted within a time frame of 30–40 minutes. We assumed this time frame to be sufficiently short for participants not losing interest. As results show, this was the case for the majority of participants. However, some participants took far more time, as can also be seen. Taking the respective participants' results for effectiveness into account, we assume that this is not related to idling (which would have made exclusion of the data sets necessary), but from participants trying to show their best performance on the study. While this is a threat for comparing different inspection techniques etc. it is not in our case. Particularly, this is a good simulation for increased effort spend compared to moderate and least possible effort.

Since volunteers may bias the results because they are generally more motivated than the average student, we decided to conduct the experiments as a mandatory part of our requirements engineering courses and explicitly decided to give no bonuses or credits as motivation. Therefore, the experiments were designed to also serve as teaching material, achieving a learning effect on model perception. This was supported by extensive debriefings in class. The experimental setup was carefully adopted to meet national laws as well as comply with university's ethics

(a) Results of Games-Howell-Test for Confidence

		MD	SE	Sig.	95% Conf. Interval	
1-5min	*5-10min*	-0.577	0.114	0.000	-0.926	-0.229
	10-15min	-1.055	0.110	0.000	-1.392	-0.718
	15-20min	-1.082	0.118	0.000	-1.445	-0.718
	20-25min	-0.755	0.246	0.106	-1.616	0.105
	25-30min	-1.469	0.236	0.002	-2.352	-0.587
	30-35min	-1.475	0.249	0.011	-2.544	-0.407
	>35min	0.590	0.447	0.865	-1.304	2.483
5-10min	*1-5min*	0.577	0.114	0.000	0.229	0.926
	10-15min	-0.478	0.114	0.001	-0.825	-0.131
	15-20min	-0.504	0.121	0.001	-0.877	-0.132
	20-25min	-0.178	0.247	0.995	-1.041	0.685
	25-30min	-0.892	0.237	0.047	-1.775	-0.009
	30-35min	-0.898	0.251	0.103	-1.964	0.168
	>35min	1.167	0.448	0.296	-0.726	3.059
10-15min	*1-5min*	1.055	0.110	0.000	0.718	1.392
	5-10min	0.478	0.114	0.001	0.131	0.825
	15-20min	-0.026	0.117	1.000	-0.388	0.336
	20-25min	0.300	0.246	0.913	-0.560	1.160
	25-30min	-0.414	0.235	0.657	-1.296	0.468
	30-35min	-0.420	0.249	0.697	-1.489	0.649
	>35min	1.645	0.447	0.091	-0.249	3.539
15-20min	*1-5min*	1.082	0.118	0.000	0.718	1.445
	5-10min	0.504	0.121	0.001	0.132	0.877
	10-15min	0.026	0.117	1.000	-0.336	0.388
	20-25min	0.326	0.249	0.882	-0.540	1.192
	25-30min	-0.388	0.239	0.730	-1.273	0.497
	30-35min	-0.394	0.252	0.759	-1.458	0.671
	>35min	1.671	0.449	0.085	-0.220	3.562
20-25min	*1-5min*	0.755	0.246	0.106	-0.105	1.616
	5-10min	0.178	0.247	0.995	-0.685	1.041
	10-15min	-0.300	0.246	0.913	-1.160	0.560
	15-20min	-0.326	0.249	0.882	-1.192	0.540
	25-30min	-0.714	0.322	0.384	-1.803	0.375
	30-35min	-0.720	0.332	0.423	-1.892	0.453
	>35min	1.345	0.498	0.230	-0.546	3.236
25-30min	*1-5min*	1.469	0.236	0.002	0.587	2.352
	5-10min	0.892	0.237	0.047	0.009	1.775
	10-15min	0.414	0.235	0.657	-0.468	1.296
	15-20min	0.388	0.239	0.730	-0.497	1.273
	20-25min	0.714	0.322	0.384	-0.375	1.803
	30-35min	-0.006	0.325	1.000	-1.182	1.170
	>35min	2.059	0.493	0.032	0.166	3.952
30-35min	*1-5min*	1.475	0.249	0.011	0.407	2.544
	5-10min	0.898	0.251	0.103	-0.168	1.964
	10-15min	0.420	0.249	0.697	-0.649	1.489
	15-20min	0.394	0.252	0.759	-0.671	1.458
	20-25min	0.720	0.332	0.423	-0.453	1.892
	25-30min	0.006	0.325	1.000	-1.170	1.182
	>35min	2.065	0.500	0.033	0.149	3.981
>35min	*1-5min*	-0.590	0.447	0.865	-2.483	1.304
	5-10min	-1.167	0.448	0.296	-3.059	0.726
	10-15min	-1.645	0.447	0.091	-3.539	0.249
	15-20min	-1.671	0.449	0.085	-3.562	0.220
	20-25min	-1.345	0.498	0.230	-3.236	0.546
	25-30min	-2.059	0.493	0.032	-3.952	-0.166
	30-35min	-2.065	0.500	0.033	-3.981	-0.149

(b) **Results of Games-Howell-Test for Efficiency**

		MD	SE	Sig.	95% Conf. Interval	
1-5 min	*5-10min*	-9.678	5.213	0.582	-25.638	6.283
	10-15min	7.262	6.612	0.956	-13.113	27.636
	15-20min	5.342	11.927	1.000	-32.536	43.219
	20-25min	14.540	15.646	0.977	-41.759	70.839
	25-30min	-6.806	37.197	1.000	-153.765	140.153
	30-35min	-132.707	74.154	0.650	-477.626	212.212
	>35min	-747.501	166.194	0.042	-1466.889	-28.113
5-10 min	*1-5min*	9.678	5.213	0.582	-6.283	25.638
	10-15min	16.940	7.924	0.394	-7.313	41.192
	15-20min	15.019	12.701	0.934	-24.911	54.949
	20-25min	24.217	16.244	0.801	-32.856	81.291
	25-30min	2.871	37.452	1.000	-144.035	149.778
	30-35min	-123.029	74.282	0.713	-467.509	221.450
	>35min	-737.823	166.252	0.045	-1457.102	-18.545
10-15 min	*1-5min*	-7.262	6.612	0.956	-27.636	13.113
	5-10min	-16.940	7.924	0.394	-41.192	7.313
	15-20min	-1.920	13.336	1.000	-43.608	39.768
	20-25min	7.278	16.745	1.000	-50.587	65.142
	25-30min	-14.068	37.672	1.000	-160.955	132.818
	30-35min	-139.969	74.394	0.607	-484.075	204.137
	>35min	-754.763	166.302	0.040	-1473.946	-35.579
15-20 min	*1-5min*	-5.342	11.927	1.000	-43.219	32.536
	5-10min	-15.019	12.701	0.934	-54.949	24.911
	10-15min	1.920	13.336	1.000	-39.768	43.608
	20-25min	9.198	19.466	1.000	-54.598	72.994
	25-30min	-12.148	38.958	1.000	-159.353	135.058
	30-35min	-138.049	75.053	0.625	-480.081	203.983
	>35min	-752.843	166.598	0.041	-1471.472	-34.213
20-25 min	*1-5min*	-14.540	15.646	0.977	-70.839	41.759
	5-10min	-24.217	16.244	0.801	-81.291	32.856
	10-15min	-7.278	16.745	1.000	-65.142	50.587
	15-20min	-9.198	19.466	1.000	-72.994	54.598
	25-30min	-21.346	40.252	0.999	-169.903	127.211
	30-35min	-147.247	75.733	0.575	-487.449	192.956
	>35min	-762.041	166.905	0.038	-1480.118	-43.963
25-30 min	*1-5min*	6.806	37.197	1.000	-140.153	153.765
	5-10min	-2.871	37.452	1.000	-149.778	144.035
	10-15min	14.068	37.672	1.000	-132.818	160.955
	15-20min	12.148	38.958	1.000	-135.058	159.353
	20-25min	21.346	40.252	0.999	-127.211	169.903
	30-35min	-125.901	82.911	0.781	-459.546	207.745
	>35min	-740.695	170.282	0.042	-1454.028	-27.362
3-35 min	*1-5min*	132.707	74.154	0.650	-212.212	477.626
	5-10min	123.029	74.282	0.713	-221.450	467.509
	10-15min	139.969	74.394	0.607	-204.137	484.075
	15-20min	138.049	75.053	0.625	-203.983	480.081
	20-25min	147.247	75.733	0.575	-192.956	487.449
	25-30min	125.901	82.911	0.781	-207.745	459.546
	>35min	-614.794	181.965	0.102	-1329.378	99.790
>35 min	*1-5min*	747.501	166.194	0.042	28.113	1466.889
	5-10min	737.823	166.252	0.045	18.545	1457.102
	10-15min	754.763	166.302	0.040	35.579	1473.946
	15-20min	752.843	166.598	0.041	34.213	1471.472
	20-25min	762.041	166.905	0.038	43.963	1480.118
	25-30min	740.695	170.282	0.042	27.362	1454.028
	30-35min	614.794	181.965	0.102	-99.790	1329.378

Figure 3: Results of Games-Howell-Test.

regulations on student participation in software engineering experiments.

5.2.2 Threats to Construct Validity

The experiment setup was inspired by conducted, published, and well-received experiments from the related work on the investigation of different inspection techniques. The experiment material was created in close collaboration with domain experts from industry and academia. In addition, pretest groups have been used to ensure that experiment material is comprehensible and adequate for the given tasks.

5.2.3 Threats to External Validity

External validity in software engineering experiments is mainly concerned with the question of generalizability to industrial application (Höst et al., 2000). Therefore, the use of student participants is often seen as problematic (e.g., (Runeson, 2003)). However, other studies have found out that student results are generalizable (e.g., (Tichy, 2000)). In addition, ad hoc inspections are in industrial practice often conducted by newer employees, for which generalizability from students often holds (Salman et al., 2015). For the experiment material, we ensured generalizability in close collaboration with industry professionals to adopt excerpts from industry specifications.

To improve generalizability, we used models in three different modeling languages. While we, thus, do not need to limit our results to one single modeling language, there is a risk, that review time, effectiveness, confidence, efficiency significantly differ between the different modeling languages. While we did not recognize such effects, we ensured that each participant conducted ad hoc model inspection tasks for models in different modeling languages. To avoid crossover effects having an impact, we used randomization to distribute the order of inspection tasks equally over all participants.

5.2.4 Threats to Conclusion Validity

The major threat regarding conclusion validity is typically the use of too small sample sizes, which hinder reaching statistical significance. The use of 497 included data sets is to be considered large compared to other investigations from the related work. Another threat to conclusion validity lies in transforming ratio scale data into ordinal scale data, as has been done for review time to allow for conducting analysis of variance. There is the risk, that using another interval leads to different results. Therefore, we conducted a second investigation, grouping review time into intervals of one minute. Results do not considerably differ. Particularly, no new significant differences could be found (e.g., differences between two medium review time intervals). Therefore, we assume our grouping adequate.

5.3 Inferences

Most important is the insight that no empirical evidence could be found that review time has any influence on the effectiveness of ad hoc model inspections. As effectiveness best relates to the number of defects found during an inspection, we can conclude that ad hoc model inspections can be conducted in short time and increasing review time does not lead to any considerable advantage regarding the number of defects found. However, it is to note that our finding is limited to the inspection of models that are reduced in size and complexity and limited to approximately one page. Hence, we cannot state that this finding will also hold for the review of specifications with multiple models or complex models consisting of a multitude of diagrams. Therefore, further investigations would be needed. However, for our definition of review time (i.e. the reviewer chooses to spend) we assume there might be no significant influence either. In this case, also in our experiment, reviewers that are more thorough and spend more review time should have achieved better results no matter the size of the materials.

Regarding the influence of review time on confidence and efficiency, we found large review times leading to less confident decisions and low efficiency. Therefore, it seems counter intuitively better to restrict review time to a moderate amount of less than 30 minutes. In addition, review times of less than 10 minutes also lead to low confidence, although efficiency is not influenced.

In summary, review time cannot be used to estimate the quality of a review. As there are no significant differences between review times of 10–30 minutes discernible, the ad hoc model inspection can be kept brief. We assume that this is transferable to larger inspections as well, keeping the review time short but not too short maximizes the result of an ad hoc model inspection. As review time does not influence the effectiveness of the ad hoc model inspection, it can be assumed that brief visual inspections conducted in very short time are helpful in the development process and, thus, should be made use of whenever possible. In particular, the insight regarding the missing influence of review time on effectiveness allows for a multitude of potential application scenarios for ad hoc model inspections during a software development project.

6 CONCLUSION

In this paper, we reported an experiment to investigate the influence of review time on ad hoc model inspections. In the experiment we analyzed the influence of review time on effectiveness, confidence, and efficiency. The experiment was conducted with 200 participants that conducted a total of 520 ad hoc model inspections. Most important, analysis of the data sets showed that review time does not have a significant influence on the effectiveness of ad hoc model inspections. For confidence, we found a small influencing effect, for efficiency a high effect.

Analysis of variance (ANOVA) showed that review time leads to significantly different confidence and efficiency. Post hoc tests showed that a short review time of up to ten minutes negatively influences the confidence the reviewer has in the decisions made. In contrast to assumptions made in the related work, we found out that large review times also have a negative influence. For a review time greater than thirty minutes, confidence and efficiency is significantly lower as for moderate review time.

REFERENCES

Albers, K., Beck, S., Büker, M., Daun, M., MacGregor, J., Salmon, A., Weber, R., and Weyer, T. (2016). System function networks. In *Advanced Model-Based Engineering of Embedded Systems, Extensions of the SPES 2020 Methodology*, pages 119–144. Springer.

Basili, V. R., Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sørumgård, S., and Zelkowitz, M. V. (1996). The Empirical Investigation of Perspective-Based Reading. *Empirical Software Engineering*, 1(2):133–164.

Bavota, G., Gravino, C., Oliveto, R., De Lucia, A., Tortora, G., Genero, M., and Cruz-Lemus, J. A. (2011). Identifying the Weaknesses of UML Class Diagrams during Data Model Comprehension. In *Model Driven Engineering Languages and Systems*, pages 168–182, Berlin, Heidelberg. Springer.

Berling, T. and Runeson, P. (2003). Evaluation of a perspective based review method applied in an industrial setting. *IEE Proceedings - Software*, 150(3):177–184.

Boehm, B. W. (1987). Industrial software metrics top 10 list. *IEEE Software*, 4(5):84–85.

Conradi, R., Mohagheghi, P., Arif, T., Hegde, L. C., Bunde, G. A., and Pedersen, A. (2003). Object-Oriented Reading Techniques for Inspection of UML Models – An Industrial Experiment. In *ECOOP 2003 – Object-Oriented Programming*, pages 483–500, Berlin, Heidelberg. Springer.

d. Mello, R. M., Teixeira, E. N., Schots, M., Werner, C. M. L., and Travassos, G. H. (2012). Checklist-based inspection technique for feature models review. In *2012 Sixth Brazilian Symposium on Software Components, Architectures and Reuse*, pages 140–149.

Daun, M., Brings, J., Krajinski, L., and Weyer, T. (2019a). On the benefits of using dedicated models in validation processes for behavioral specifications. In *IEEE/ACM Int. Conf. on Software and System Processes*, pages 44–53. IEEE.

Daun, M., Brings, J., Obe, P. A., and Stenkova, V. (2021). Reliability of self-rated experience and confidence as predictors for students' performance in software engineering. *Empirical Software Engineering*, 26(4):80.

Daun, M., Brings, J., and Weyer, T. (2017). On the impact of the model-based representation of inconsistencies to manual reviews: Results from a controlled experiment. In *Conceptual Modeling: 36th Int. Conf.*, pages 466–473. Springer.

Daun, M., Brings, J., and Weyer, T. (2020). Do instance-level review diagrams support validation processes of cyber-physical system specifications: results from a controlled experiment. In *Int. Conf. on Software and System Processes*, pages 11–20.

Daun, M., Weyer, T., and Pohl, K. (2014). Validating the functional design of embedded systems against stakeholder intentions. In *Int. Conf. on Model-Driven Engineering and Software Development*, pages 333–339. IEEE.

Daun, M., Weyer, T., and Pohl, K. (2019b). Improving manual reviews in function-centered engineering of embedded systems using a dedicated review model. *Software and Systems Modeling*, 18(6):3421–3459.

de Alfaro, L. and Henzinger, T. A. (2001). Interface automata. *SIGSOFT Softw. Eng. Notes*, 26(5):109–120.

de Almeida, J. R., Camargo, J. B., Basseto, B. A., and Paz, S. M. (2003). Best practices in code inspection for safety-critical software. *IEEE Software*, 20(3):56–63.

Doolan, E. P. (1992). Experience with fagan's inspection method. *Software: Practice and Experience*, 22(2):173–182.

Dunsmore, A., Roper, M., and Wood, M. (2003). Practical code inspection techniques for object-oriented systems: an experimental comparison. *IEEE Software*, 20(4):21–29.

Fagan, M. E. (1976). Design and Code Inspections to Reduce Errors in Program Development. *IBM Systems Journal*, 15(3):182–211.

Fagan, M. E. (1986). Advances in Software Inspections. *IEEE Trans. Software Eng.*, 12(7):744–751.

Figl, K., Mendling, J., and Strembeck, M. (2013a). The Influence of Notational Deficiencies on Process Model Comprehension. *Journal of the Association for Information Systems*, 14(6).

Figl, K., Recker, J., and Mendling, J. (2013b). A study on the effects of routing symbol design on process model comprehension. *Decision Support Systems*, 54(2):1104–1118.

He, L. and Carver, J. C. (2006). PBR vs. checklist: a replication in the n-fold inspection context. In Travassos, G. H., Maldonado, J. C., and Wohlin, C., editors, *2006 International Symposium on Empirical Software Engineering*, pages 95–104. ACM.

Höst, M., Regnell, B., and Wohlin, C. (2000). Using Students as Subjects-A Comparative Study of Students and Professionals in Lead-Time Impact Assessment. *Empirical Software Engineering*, 5(3):201–214.

International Telecommunication Union (2016). Recommendation z.120: Message Sequence Chart (MSC).

ISO 26262-1 (2011). Road vehicles – Functional safety – Part 1: Vocabulary.

ISO/IEC 25030 (2007). Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Quality requirements.

Jedlitschka, A., Ciolkowski, M., and Pfahl, D. (2008). Reporting experiments in software engineering. In Shull, F., Singer, J., and Sjøberg, D. I. K., editors, *Guide to Advanced Empirical Software Engineering*, pages 201–228. Springer London.

Laitenberger, O. (1998). Studying the effects of code inspection and structural testing on software quality. In *International Symposium on Software Reliability Engineering*, pages 237–246.

Laitenberger, O., Atkinson, C., Schlich, M., and El Emam, K. (2000). An experimental comparison of reading techniques for defect detection in UML design documents. *Journal of Systems and Software*, 53(2):183–204.

Laitenberger, O., Emam, K. E., and Harbich, T. G. (2001). An Internally Replicated Quasi-Experimental Comparison of Checklist and Perspective-Based Reading of Code Documents. *IEEE Trans. Software Eng.*, 27(5):387–421.

Lanubile, F. and Visaggio, G. (2000). Evaluating Defect Detection Techniques for Software Requirements Inspections. page 25.

Lucia, A. D., Gravino, C., Oliveto, R., and Tortora, G. (2008). Data model comprehension: An empirical comparison of er and uml class diagrams. In *2008 16th IEEE Int. Conf. on Program Comprehension*, pages 93–102.

Maldonado, J. C., Carver, J., Shull, F., Sandra Camargo Pinto Ferraz Fabbri, Dória, E., Martimiano, L. A. F., Mendonça, M. G., and Basili, V. R. (2006). Perspective-Based Reading: A Replicated Experiment Focused on Individual Reviewer Effectiveness. *Empirical Software Engineering*, 11(1):119–142.

Martin, J. and Tsai, W. T. (1990). N-fold inspection: A requirements analysis technique. *Commun. ACM*, 33(2):225–232.

Mendling, J., Strembeck, M., and Recker, J. (2012). Factors of process model comprehension—Findings from a series of experiments. *Decision Support Systems*, 53(1):195–206.

Miller, J., Wood, M., and Roper, M. (1998). Further Experiences with Scenarios and Checklists. *Empirical Software Engineering*, 3(1):37–64.

Nugroho, A. (2009). Level of detail in UML models and its impact on model comprehension: A controlled experiment. *Information and Software Technology*, 51(12):1670–1685.

O.Oladele, R. and O. Adedayo, H. (2014). On Empirical Comparison of Checklist-based Reading and Adhoc Reading for Code Inspection. *International Journal of Computer Applications*, 87(1):35–39.

Porter, A. A., Siy, H. P., Toman, C. A., and Votta, L. G. (1997). An experiment to assess the cost-benefits of code inspections in large scale software development. *IEEE Transactions on Software Engineering*, 23(6):329–346.

Porter, A. A. and Votta, L. G. (1998). Comparing Detection Methods For Software Requirements Inspections: A Replication Using Professional Subjects. *Empirical Software Engineering*, 3(4):355–379.

Porter, A. A., Votta, L. G., and Basili, V. R. (1995). Comparing Detection Methods for Software Requirements Inspections: A Replicated Experiment. *IEEE Trans. Software Eng.*, 21(6):563–575.

Regnell, B., Runeson, P., and Thelin, T. (2000). Are the Perspectives Really Different? – Further Experimentation on Scenario-Based Reading of Requirements. *Empirical Software Engineering*, 5(4):331–356.

Runeson, P. (2003). Using Students as Experiment Subjects – An Analysis on Graduate and Freshmen Student Data. In *International Conference on Empirical Assessment & Evaluation in Software Engineering*, pages 95–102.

Sabaliauskaite, G., Kusumoto, S., and Inoue, K. (2004). Assessing defect detection performance of interacting teams in object-oriented design inspection. *Information & Software Technology*, 46(13):875–886.

Salman, I., Misirli, A. T., and Juristo, N. (2015). Are students representatives of professionals in software engineering experiments? In *IEEE/ACM International Conference on Software Engineering*, volume 1, pages 666–676.

Shull, F., Rus, I., and Basili, V. R. (2000). How Perspective-Based Reading Can Improve Requirements Inspections. *IEEE Computer*, 33(7):73–79.

Thelin, T., Runeson, P., and Wohlin, C. (2003). An experimental comparison of usage-based and checklist-based reading. *IEEE Transactions on Software Engineering*, 29(8):687–704.

Tichy, W. F. (2000). Hints for Reviewing Empirical Work in Software Engineering. *Empirical Software Engineering*, 5(4):309–312.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M., Regnell, B., and Wesslén, A. (2000). *Experimentation in software engineering: An introduction*, volume 6 of *Kluwer international series in software engineering*. Kluwer Academic, Boston, Mass.

Zimoch, M., Pryss, R., Probst, T., Schlee, W., and Reichert, M. (2017). Cognitive Insights into Business Process Model Comprehension: Preliminary Results for Experienced and Inexperienced Individuals. In *Enterprise, Business-Process and Information Systems Modeling*, pages 137–152, Cham. Springer.