

Assessing the Frequency of Empirical Evaluation in Software Modeling Research

Jeffrey C. Carver, Eugene Syriani, Jeff Gray

University of Alabama, Department of Computer Science
Tuscaloosa, Alabama USA
{carver, esyriani, gray}@cs.ua.edu

Abstract. Researchers in software modeling often publish new tools or methodologies that claim to offer some advantage to the modeling community. There are different methods by which those claims can be evaluated. In this paper, we examine the degree to which such claims are supported by various types of empirical evaluation. We surveyed five editions of the MoDELS conference from 2006-2010, as well as the primary conference that focuses on empirical software engineering (the *International Symposium on Empirical Software Engineering and Metrics*), to understand the frequency with which empirical evaluation has been reported in the software modeling community. Our summary of 266 MoDELS papers found that 195 (73%) of the publications performed no empirical evaluation. This paper summarizes our findings from that survey and offers recommendations for improving the awareness and need for empirical evaluation in software modeling research.

Keywords: Empirical software engineering, Model-driven Engineering

1. Introduction

Research into software modeling has attracted many creative and transformative ideas over the past decade, ranging from new methods for defining languages and transforming their model instances, to higher level performance analysis and verification tools that abstract the essence of some system property. Although the novelty of software modeling research has led to numerous advances, the collective body of work in this area has not always followed the typical tenets of a scientific discipline. One of the key precepts of scientific investigation is the ability to repeat an experiment to verify that some new scientific discovery can be confirmed under numerous scenarios. For most contributions in model-driven engineering, some new tool or technique is often proposed and discussed through an illustrative case study, but generally is not evaluated at the level of rigor assumed for a traditional empirical evaluation.

Our suspicion about the level of empirical studies in modeling research led to this summary paper that analyzes the degree of empirical evaluation in software modeling research. To approach this topic, we analyzed the most recent five editions (from 2006 through 2010) of the most influential conference in software modeling – the

conference on *Model-Driven Engineering, Languages and Systems* (MoDELS). Two of the authors of this paper (Gray and Syriani) have themselves published papers at this conference that did not contain an empirical evaluation. We were curious about the extent to which this practice is common in the software modeling community. In addition to observing contributions at MoDELS, we also considered the prevalence of modeling papers at a venue focused on empirical software engineering. The remainder of this paper summarizes our findings from an analysis of 266 MoDELS papers. Our suspicions were confirmed by our analysis, which suggests that a large majority of research papers in the modeling community fail to provide any level of empirical evidence to support the claims of benefit made in those papers.

The next section of this paper provides an overview of empirical studies and the methodology that we used in conducting our analysis of MoDELS papers. Section 3 presents the results of our analysis of the MoDELS conference and our analysis of software modeling papers that have appeared in the flagship empirical software engineering conference, the *International Symposium on Empirical Software Engineering and Measurement* (ESEM). Section 4 discusses the results of the analysis in more detail. Finally, we conclude the paper in Section 5.

2. Overview of Empirical Studies and Methodology

For a new modeling tool or technique to become used, the developer of the tool or technique must demonstrate its value. Although a proof-of-concept or illustrative example are important first steps in establishing the usefulness of a technique or tool, claims about the usefulness of modeling techniques and tools cannot be fully validated without the use of various types of empirical studies. An empirical study is a validation method that draws conclusions based on observations (as opposed to proof, argumentation, or expert opinion).

In the larger software engineering community, empirical studies have commonly been used to understand developer behavior in a number of important areas. There is an entire sub-community focused on validating software engineering claims via empirical study. This sub-community has a conference (ESEM), a Springer journal (*Empirical Software Engineering*) and a number of handbooks [3], [9], [14]. The first author of this paper comes from this community.

The goal of this investigation was to determine how many papers had some type of empirical evaluation of their claims. We realize that there are evaluation methods other than empirical studies (e.g., demonstration/proof-of-concept or theoretical proof). But, in this paper, we focus only on empirical evaluation. Among the three authors, two are experts in the modeling domain and one is an expert in the empirical software engineering domain. Working together, we were able to complement each other's expertise to perform this analysis.

We used a three-step process for identifying which papers contain an empirical component. The first step was to develop an initial characterization scheme. Next, the two modeling experts individually analyzed the proceedings of various years of the MoDELS proceedings to identify and classify the papers. Third, the empirical studies expert reviewed the papers identified in step two and validated the classification of

those papers. Step 3 resulted in some modifications to the characterization scheme. The remainder of this section describes each step in more detail.

2.1. Step 1 – Develop an initial characterization scheme

We began with the assumption that there are two types of empirical studies: those that are more analytical (i.e., perform some type of analysis of a tool and its properties without using humans) and those that are human-based (i.e., they involved studying one or more people using a modeling technique). For each type of study, we created two categories: “non-rigorous” and “rigorous.” The difference between rigorous and non-rigorous was subjective and ill-defined at this first stage of analysis.

Because we had no preconceived notions of the results of the literature search, this initial characterization scheme was necessarily vague. We realized that after examining the actual papers, we would have to refine the characterization scheme to accurately describe the identified papers.

2.2. Step 2 – Identification of candidate papers

The two modeling experts divided the five years of MoDELS proceedings between them and individually analyzed all of the papers. For each paper, they first determined whether there was any type of empirical study and whether it was human-based. At this stage, they also made a subjective determination as to whether a paper was rigorous. After this step, we developed a spreadsheet that characterized each paper into one of five categories: *no empirical study*, *non-rigorous non-human*, *rigorous non-human*, *non-rigorous human* and *rigorous human*.

2.3. Step 3 – Review of candidate papers and finalization of characterization

The empirical software engineering expert then reviewed each paper that the modeling experts identified during Step 2 as having an empirical study. The goal of this process was to provide a second observation to validate the characterization from Step 2. During the review, it quickly became apparent that our initial characterization scheme was inadequate. We refined the initial characterization as follows.

First, we more clearly defined the term “empirical study.” Some of the candidate papers identified during Step 2 really contained just a demonstration or implementation of the new tool or technique rather than an empirical study. In fact, several MoDELS papers had an “Evaluation” section that was merely a discussion of lessons learned, rather than what those in the empirical software engineering community would call an empirical study. We clarified the definition of what we considered as an empirical study to exclude papers that clearly did not gather any type of data to evaluate the proposed tool or technique.

In reviewing the papers, we identified two types of empirical papers:

1. Papers that propose a new tool or technique and then perform some type of evaluation of it.
2. Papers that gather information about the use of modeling techniques in practice. These papers do not propose new approaches; rather, they study existing approaches or survey users to develop requirements for tools or techniques that may be needed. We call these papers Formative Case Studies, as opposed to the Illustrative Case Studies that just illustrate the use of a new tool or technique.

Second, we refined the original characterization scheme to define more concretely the categories into which the papers could be classified. The revised characterization scheme is as follows:

1. *No empirical evaluation* – the paper did not provide any type of empirical evaluation of the proposed tool or technique (this, unfortunately, represented the overwhelming majority of the papers we analyzed).
2. *Non-human evaluation of the proposed tool/technique only* – the paper offered some type of empirical evaluation (e.g., performance or correctness) of the proposed tool, but did not compare the new tool against other tools or benchmarks.
3. *Non-human evaluation of proposed tool/technique by comparison with other tools* – the paper provided an empirical evaluation by comparing the proposed tool/technique against one or more existing tools or benchmarks to evaluate some aspect of the new tool/technique.
4. *Observation of humans using new tool/technique* – the paper discussed and analyzed the results from the use of the new tool/technique by one or more people other than the authors of the paper.
5. *Human-based controlled experiment* – the paper described a controlled experiment where the new tool/technique was compared against one or more existing approaches through a human-based controlled experiment where each participant used one or more approaches and provided data that could be analyzed to evaluate the new tool/technique.
6. *Formative case study* – as defined above.

3. Results of Literature Survey

This section summarizes the results of our survey of the MoDELS papers and of the modeling papers that appeared in the ESEM conference.

Table 1. Results of the survey of the papers published at MoDELS 2006-2010

Year	Total	No Evaluation	Non-Human		Human		
			No comparison	Comparison	Observation	Controlled Experiment	Formative Case Study
2006	51	42 (82%)	6 (12%)	0 (0%)	1 (2%)	1 (2%)	1 (2%)
2007	45	36 (80%)	2 (4%)	5 (11%)	0 (0%)	2 (4%)	0 (0%)
2008	58	39 (67%)	8 (14%)	2 (3%)	2 (3%)	4 (7%)	3 (5%)
2009	58	45 (78%)	5 (9%)	2 (3%)	2 (3%)	1 (2%)	3 (5%)
2010	54	33 (61%)	8 (15%)	4 (7%)	2 (4%)	4 (7%)	3 (6%)
Total	266	195 (73%)	29 (10%)	13 (4%)	7 (2%)	12 (4%)	10 (3%)

3.1. Results of MoDELS Survey

In the empirical evaluations conducted, we analyzed a total of 266 papers published at MoDELS from 2006-2010. The complete analysis of the papers took approximately 18 hours of observation and recording. Table 1 summarizes the results of this assessment.

It is very clear that, for each year, the number of papers without any evaluation was predominant: ranging from 61% in 2010 up to 82% in 2006. However, the tendency seems to suggest a rising awareness and influence of the need for empirical studies, as we note an average decrease of about 4% each year in the number of papers with no evaluation (there is a 21% drop in the “No Evaluation” category from the beginning of our study period to the end of the period over the five years observed). We have no direct evidence for the cause of this improvement, but feedback sent to authors on reviews over the period of the study may suggest the emerging demand among the Program Committee for more rigorous evaluation.

Those papers that did have some form of empirical study were often restricted to simple evaluations of performance or correctness of the proposed tool/technique without comparing it to other results (41% of the those papers describing an empirical study were in the “No comparison” category). The papers in 2007 seem to be the only exception, where 11% of all papers addressed comparisons with other tools or benchmarks.

On average, about 11% of the papers were supported by empirical studies involving humans. In this category, 42% of the papers contained controlled experiments, representing not more than 7% of all papers (years 2008 and 2010). The number of papers where the evaluation was observed by at least one external participant has been quite steady at about 3% of all papers. Formative case studies are gaining popularity with up to 6% of all the papers in 2010.

The “Total” row (at the bottom of Table 1) shows the portions occupied by each of the categorizations defined in Section 2 across all years. Although 73% of the papers published at MoDELS do not contain an evaluation, 10% of the papers only evaluate their own tool without any comparison to other approaches. Thus, only the remaining 17% of the papers involve an empirical evaluation of the proposed tool or technique. However, according to Fig. 1, this number is increasing every year: up to 24% in 2010. This trend may suggest that authors are aware of the lack of empirical evidence in the modeling community and are now working on filling this gap.

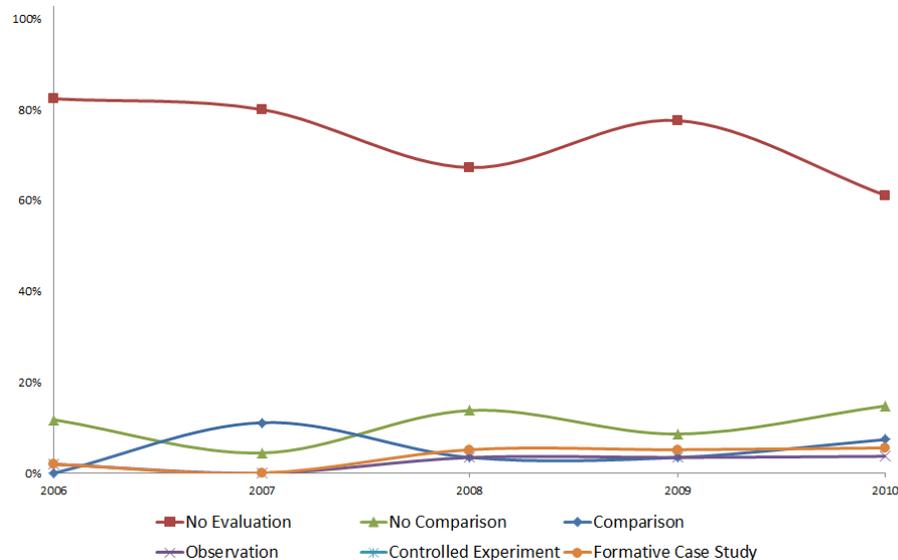


Fig. 1. Evolution of the empirical studies involved in MoDELS 2006-2010 papers

3.2 Results of ESEM Survey

As evidenced by the discussion in the previous section, the MoDELS conference appears to be focused mainly on proposing new tools and techniques without rigorous evaluation. To the authors' credit, the paper length restrictions of the LNCS format used in MoDELS leave little space for discussion of formal evaluation. The *ESEM* conference is a general software engineering conference that focuses on the empirical evaluation of newly proposed techniques across all software engineering topics. We analyzed the same five years of the ESEM conference to determine whether more formal evaluations of modeling research were being published there. To identify the set of papers, we queried the proceedings using the following keywords: "UML," "DSL," "metamodel," "model," and "model-driven." The modeling experts then vetted the results of the search to ensure that the papers were within the scope of software modeling.

Based on this analysis, we can make a few interesting observations. The ESEM conference has three types of papers: Regular Papers, Short Papers, and Posters. In total, we only found 17 modeling papers across the five years that we analyzed ESEM. Of those 17 papers, only 4 were Regular Papers (10 pages IEEE or ACM format) out of a total of 178 Regular Papers and 10 were Short Papers (4 pages) out of a total of 118 Short Papers. Thus, even when software modeling papers are published in an empirical venue, they tend to be shorter and do not provide a high-level of detail. In analyzing the five years of ESEM, we were not able to identify any trends

that would suggest the prominence of modeling papers is increasing in the empirical software engineering community. One final observation, in comparing the author lists and titles of the ESEM papers against the empirical MoDELS papers, we found very little overlap; only one paper seemed to be about the same tool or technique. Thus, the cross-pollination of results across the two communities seems to be very low.

4. Observations from Our Survey

This section provides a summary of our observations about the papers that focused on controlled experiments and formative case studies.

4.1. Controlled Experiments

Across the five years of the MoDELS conference, we found twelve controlled experiments [1], [2], [5], [6], [7], [8], [10], [11], [12], [13], [15], [16]. This category of papers serves as an example of the types of papers that we feel should be more prevalent within the MoDELS community. In this section, we provide a brief discussion of some of the trends observed in these controlled experiment papers. Overall, the level of detail reported by the authors of these papers is quite low. We realize that this level of reporting is likely affected by paper length restrictions and the need to fully describe the newly proposed tool or technique as the core contribution of the paper. Although we do not have the space to evaluate the quality of each study in detail, there are two important factors that are relatively easy to evaluate: 1) the number of participants, and 2) whether the participants were students or professionals.

In terms of the number of participants in the studies, one half of the identified papers had less than 25 participants, and only two studies had more than 50 participants. Furthermore, one study did not even report the number of participants. In terms of the type of participant, only one study had professionals as a portion of the participants. The overwhelming majority of the studies relied on undergraduates with only a few using graduate students. Over 33% of the studies did not specify whether the participants were students or professionals. The use of student participants is not necessarily bad, but researchers need to make a clear case as to why student participants are a valid population for the question under investigation [4].

There does not appear to be a significant trend in the number of controlled experiments reported. From 2006 through 2008, the number was increasing. Then, there was a large drop of such experiments in 2009. The percentage of controlled experiments in 2010 was equivalent to the percentage reported in 2008. Even in the best years, only 7% of the papers reported controlled experiments. In general, we would like to see an increase in both the frequency and diversity of controlled experiments within the modeling community.

4.2. Formative Case Studies

Across the five years of the MoDELS conference, we found ten Formative Case Study papers. There were two types of Formative Case Studies. First, there were four studies that did not involve humans. These studies tended to analyze some existing source code to understand how various modeling tools would or would not work effectively. Second, there were six studies that focused on humans. These studies mostly used a survey method to understand how existing tools were not meeting the needs of developers. The output of many of these studies was a set of requirements for new tools that were needed. Contrary to the Controlled Experiments, which focused heavily on student participants, the Formative Case Studies were focused more on industrial settings. Similar to the Controlled Experiments, we would also like to see additional Formative Case Studies that provide input to tool and method developers to help ensure that their work is relevant to the needs of practitioners.

5. Conclusion

This paper provides evidence that the rigor of empirically validated research in software modeling is rather weak and should be a focus of future authors of MoDELS papers. The high-level of incidence of papers with no evaluation is somewhat alarming when compared to other software engineering venues (e.g., ICSE) where empirical evaluation is more expected as a scientific contribution. Overall, the level of empirical evaluation as seen in the software modeling community is quite low for a scientific and engineering discipline. A goal of this paper is to raise the awareness of this issue to assist in progressing the area of software modeling with a more scientific underpinning. Our own future work will include a similar analysis of papers in the software modeling community's flagship journal – *Software and Systems Modeling*.

As part of this work, we posit that there is a need for more controlled experiments within the modeling community. We realize that there are at least three factors that are hindering these types of studies being conducted. First, many researchers in the modeling community may lack the background or training to carry out empirical studies. This situation is evidenced by the fact that authors frequently mention “validation experiments” which are nothing more than the application of the findings or a toy example. Second, many researchers in the modeling community are more interested in creating new tools and techniques than they are in performing the rigorous evaluation of those techniques. Third, given the length restrictions of the formatting style in the MoDELS conference, there is often not adequate space to discuss both the new tool or technique and its validation, so most researchers seem to opt for devoting space to the definition of the tool or technique as representing the core contribution of their paper.

Our goal in this paper is to stress the importance of building a culture that values and expects empirical validation of newly proposed tools and methods. To help facilitate this goal, we propose the following solutions to the problem. First, researchers in the modeling domain who are interested in conducting appropriate empirical evaluations themselves need to collaborate more often with researchers who

have expertise in empirical evaluation of software engineering methods (as the authors of this paper are doing). Such a collaboration allows both types of researchers to do what they are interested in and what they do best. Second, we suggest that more rigorous empirical evaluations of modeling research be published in the ESEM conference, where the focus is on the empirical evaluation, to cross-pollinate the contributions of the modeling community with those explicitly working in empirical techniques. In that venue, authors can devote more space to describing the evaluation and interpreting the results. A somewhat radical suggestion is to afford MoDELS authors an additional two to three pages of space for any paper that includes a more rigorous evaluation based on an empirical study.

A spreadsheet representing the results of our analysis of MoDELS conferences, and a summary of the papers analyzed for the ESEM conferences, is available at: <http://www.cs.ua.edu/~carver/Data/2011/EESMOD/>

Acknowledgments. This research was supported in part by NSF CAREER award CCF-1052616.

References

1. Almeida da Silva, M., Bendraou, R., Blanc, X. et al.: Early Deviation Detection in Modeling Activities of MDE Processes. In: Petriu, D., Rouquette, N. and Haugen, Ø. (eds.) Model Driven Engineering Languages and Systems, LNCS vol. 6395, pp. 303-317. Oslo, Norway (2010)
2. Almeida da Silva, M., Mougnot, A., Bendraou, R. et al.: Artifact or Process Guidance, an Empirical Study. In: Petriu, D., Rouquette, N. and Haugen, Ø. (eds.) Model Driven Engineering Languages and Systems, LNCS vol. 6395, pp. 318-330. Oslo, Norway (2010)
3. Boehm, B., Rombach, H. D., Zelkowitz, M. V.: Foundations of Empirical Software Engineering: The Legacy of Victor R. Basili. Springer (2005)
4. Carver, J., Jaccheri, L., Morasca, S. et al.: A Checklist for Integrating Student Empirical Studies with Research and Teaching Goals. *Empirical Software Engineering*, **15** (2010) 35-59
5. Correa, A., Werner, C., Barros, M.: An Empirical Study of the Impact of OCL Smells and Refactorings on the Understandability of OCL Specifications. In: Engels, G., Opdyke, B., Schmidt, D., et al (eds.) Model Driven Engineering Languages and Systems, LNCS vol. 4735, pp. 76-90. Nashville, TN (2007)
6. Fuhrmann, H., & von Hanxleden, R.: Taming Graphical Modeling. In: Petriu, D., Rouquette, N. and Haugen, Ø. (eds.) Model Driven Engineering Languages and Systems, LNCS vol. 6394, pp. 196-210. Oslo, Norway (2010)
7. Genero, M., Cruz-Lemus, J., Caivano, D. et al.: Assessing the Influence of Stereotypes on the Comprehension of UML Sequence Diagrams: A Controlled Experiment. In: Czarnecki, K., Ober, I., Bruel, J., et al (eds.) Model Driven Engineering Languages and Systems, LNCS vol. 5301, pp. 280-294. Toulouse, France (2008)

8. Gravino, C., Scanniello, G., Tortora, G.: An Empirical Investigation on Dynamic Modeling in Requirements Engineering. In: Czarnecki, K., Ober, I., Bruel, J., et al (eds.) Model Driven Engineering Languages and Systems, LNCS vol. 5301, pp. 615-629. Toulouse, France (2008)
9. Juristo, N., & Moreno, A.: Lecture notes on empirical software engineering. World Scientific, Singapore (2003)
10. Lange, C., DuBois, B., Chaudron, M. et al.: An Experimental Investigation of UML Modeling Conventions. In: Nierstrasz, O., Whittle, J., Harel, D., et al (eds.) Model Driven Engineering Languages and Systems, LNCS vol. 4199, pp. 27-41. Genova, Italy (2006)
11. Lucrédio, D., de M. Fortes, R., Whittle, J.: MOOGLE: A Model Search Engine. In: Czarnecki, K., Ober, I., Bruel, J., et al (eds.) Model Driven Engineering Languages and Systems, LNCS vol. 5301, pp. 296-310. Toulouse, France (2008)
12. Mäder, P., & Cleland-Huang, J.: A Visual Traceability Modeling Language. In: Petriu, D., Rouquette, N. and Haugen, Ø. (eds.) Model Driven Engineering Languages and Systems, LNCS vol. 6394, pp. 226-240. Oslo, Norway (2010)
13. Prochnow, S., & von Hanxleden, R.: Statechart Development Beyond WYSIWYG. In: Engels, G., Opdyke, B., Schmidt, D., et al (eds.) Model Driven Engineering Languages and Systems, LNCS vol. 4735, pp. 635-649. Nashville, TN (2007)
14. Shull, F., Singer, J., Sjøberg, D. I. K.: Guide to Advanced Empirical Software Engineering. Springer (2008)
15. Stålhane, T., & Sindre, G.: Safety Hazard Identification by Misuse Cases: Experimental Comparison of Text and Diagrams. In: Czarnecki, K., Ober, I., Bruel, J., et al (eds.) Model Driven Engineering Languages and Systems, LNCS vol. 5301, pp. 721-735. Toulouse, France (2008)
16. Yue, T., Briand, L., Labiche, Y.: A Use Case Modeling Approach to Facilitate the Transition towards Analysis Models: Concepts and Empirical Evaluation. In: Schürr, A. and Selic, B. (eds.) Model Driven Engineering Languages and Systems, LNCS vol. 5795, pp. 484-498. Denver, CO (2009)