On Evaluating Multi-Level Modeling

Colin Atkinson[1] and Thomas Kühne[2]

[1] University of Mannheim,
[2] Victoria University of Wellington,

Abstract. Multi-Level Modeling is receiving increasing levels of interest and its active research community is continuing to make progress. However, to advance the discipline effectively it is necessary to increase industry adoption and achieve better community cohesion. We believe that the key to addressing both these challenges is to promote the creation of more comparisons in the multi-level modeling field based on meaningful objective evaluations. In this position paper, we provide our view on what constitutes *meaningful evaluations* and discuss some of the issues involved in obtaining them, while presenting a broad overview of existing multi-level modeling evaluations. In particular, we emphasize the importance of understanding and managing the difference between *internal* and *external* qualities.

1 Introduction

Although Multi-level Modeling (MLM) has seen steady development over recent years, industry adoption is still virtually non-existent (a rare application of MLM in an industry setting is described in [8]). One explanation for the low adoption rate is the current unavailability of industrial-strength approaches and tools. However, even if better tool support were available, wider adoption would still be hindered by the lack of compelling evidence that switching from Two-Level Modeling (TLM) to MLM brings benefits in industrial contexts. Creating convincing comparisons would reduce this barrier and could even expedite MLM research through industrial funding.

Another obstacle to the discipline's future growth is the lack of research cohesion that may eventually make it impossible for community members to build on each other's results. Without a sense of direction — i.e., a common understanding of the way forward – the discipline runs the risk of research diversification to a point where it loses its core focus and subsequently critical mass. We therefore believe that objective comparisons between competing approaches are not only desirable to provide a compass for future development but may eventually become necessary for the discipline's survival.

Since convincing comparative evaluations are the key to addressing both of the aforementioned challenges, in this position paper we discuss some of the issues involved in performing such evaluations in the context of MLM. We first establish the basic parameters of meaningful, scientifically sound evaluations and then discuss a variety of concrete approaches, pointing to existing work where applicable.

2 Meaningful Evaluations

The effectiveness of MLM evaluations in promoting industry adoption and research cohesion depends on the extent to which they measure something of relevance. Naturally, relevance itself depends on the stakeholders and their respective goals. However, in general, a *meaningful evaluation* provides results that have some kind of real-world relevance. In contrast, a meaningless evaluation – e.g., measuring the number of vowels in a language's keywords – has no such real-world relevance. A comparison based on such an evaluation would not yield any meaningful insights into which language should be preferred for achieving any reasonable real-world impact. Meaningful evaluations, on the other hand, should be designed to deliver some insights that provide the basis for pragmatic guidance. In order for an evaluation to be meaningful in our sense, it must address the following aspects:

A_1: Measurability Any targeted properties must be objectively observable. Ideally, measurements should yield numeric results that are directly proportional to the property being measured, as it is then not only possible to decide which approach is better but also by how much. It is never possible to judge an approach or tool to be, e.g., "good" or "productive", without breaking down how the quality concerned manifests itself in terms of measurable properties.

A_2: Conclusiveness Measurements should yield consistent results. Repeated performances of the evaluation need not yield exactly the same outcome, but they should deliver reliable values within a given margin of error. This also excludes results with low confidence (e.g., because they lack statistical significance).

A_3: Impartiality The choice of the properties to be measured must not favor particularities of one solution that have no proven relationship to the ultimate goal. For example, a set of postulated requirements must be formulated in such a manner that they reference the problem-domain and the ultimate benefits to the targeted user rather than solution details.

A_4: Trueness When using proxies (e.g. substitutes for real-world artifacts or practitioners) care must be taken to ensure that no circumstantial bias is introduced. Trueness therefore comprises at least:

$A_{4.1}$: Context Relevance Model proxies (i.e., samples used in lieu of real-world models) and the assumed operations on them should be demonstrated to be representative. Otherwise, a skewed selection could introduce undesired bias.

$A_{4.2}$: Demographic Relevance Substitute users should be demonstrated to be representative of real users. In general, it is not possible to transfer results between different bodies of users (e.g., from students to practitioners in the field).

A_5: Pragmatic Relevance Targeted properties must have a bearing on the actual needs of the intended users. This criterion is the very foundation of a *meaningful* evaluation. The previous aspects essentially characterize *sound* evaluations, whereas *pragmatic relevance* requires that there is an intent to measure something of pragmatic value.

It is obviously challenging to "tick" all the above "boxes" in practice, but we feel it is useful to have a checklist that helps to document where an evaluation may be lacking.

3 Internal versus External Qualities

Some of the aforementioned aspects are more difficult to address than others. In order to understand why, it is important to be aware of whether an evaluation is intended to evaluate an *internal quality* or an *external quality*. We use these terms with their usual meaning in software engineering [17].

In our context, internal qualities pertain to the directly measurable properties of a model, e.g., number of model elements, number of constraints, average inheritance depth, etc. External qualities, on the other hand, pertain to the experience users have when working with a model, e.g., creating it, understanding it, maintaining it, etc.

Ultimately, only the external qualities have a direct bearing on meaningful evaluations. However, due to the cost and challenges involved in assessing external qualities directly in a meaningful way, one often attempts to approximate the assessment of external qualities by assessing internal qualities instead, based on the idea that there is a correlation between internal and external qualities. It is standard practice to assume that optimizing certain internal qualities (e.g., reducing complexity) is the key to achieving certain desirable external qualities (e.g., increased maintainability). However, such an indirect evaluation of external properties is only trustworthy if the assumed underlying correlation has been demonstrated, or at least has been made plausible by compelling arguments.

Interestingly, A_1&A_2 are most easily addressed by focusing on the internal qualities of an approach. Such qualities, e.g., the complexity of the models created by an approach, can typically be reliably assessed. In contrast, assessing external qualities often implies some compromise in A_1&A_2 because sample populations may be small or certain assumptions may not generalize.

Aspects A_3 & A_5, on the other hand, are best addressed by focusing on the external qualities of an approach. External qualities directly reflect the utility of the approach to its users and hence avoid solution bias (A_3) plus intrinsically imply pragmatic relevance (A_5). The increased cost involved in directly assessing external qualities relates to ensuring conclusiveness (A_2) and trueness (A_4). This cost is considerable and therefore represents a major hurdle for this kind of evaluation.

4 Assessing Internal Qualities

Complexity is one of the most commonly measured internal qualities since it is assumed to have a correlation with important external qualities such as maintainability, robustness, and trustworthiness etc. In fact, the main value proposition for MLM is its ability to reduce accidental complexity [3], i.e., the difference in complexity between an ideal model and a concrete model involving solution-induced overhead, e.g. workarounds.

A number of evaluations of multi-level modeling have been based on approximating the complexity of a model by measuring its size, that is, the number of its elements. For example, Gerbig performed a comparison based on model size in his Ph.D. thesis using a sample model from the enterprise architecture domain [7]. The MLM version of the model has 50 modeling elements while the TLM version, using standard workaround patterns such as the Type-Object pattern [10], has 95 modeling elements, amounting to an increase of 90%. Rossini et al. performed a similar evaluation which yielded a three-fold increase in the number of modeling elements in a two-level versus a multi-level model of their CloudML scenario [16].

The extent of the practical relevance of the above evaluations was shown by de Lara et. al. by measuring the application frequency of TLM workaround techniques (cf. "Item Descriptor" pattern [6], "Type-Object" pattern [10], "Adaptive Object-Model" [18], etc.) in real-world models [12]. Since these workaround techniques are responsible for increases of the size of two-level models relative to their multi-level counterparts, de Lara et al. hence demonstrated that the observations made in [7, 16] apply to a wide range of modeling practice. As much as 35% of all models in some areas [12], could thus benefit from the potential size reductions.

Although the above results provide a convincing endorsement for the practical relevance of MLM, they do so only to the extent that the assumption that model size[1] approximates model complexity is reasonable. A larger model based on a simple underlying language could conceivably be preferable to a compact model based on a complex language.

Going beyond assessing model size, it appears useful to consider other classic metrics [5, 13, 15] and quality attributes [4, 14]. Indeed, in his MLM vs TLM comparison, Gerbig also considered such classic metrics [7]. Overall, however, these proved to be less conclusive than model size comparisons, although he detected clear advantages for MLM with respect to coupling (*average number of distinct connected classes*) and overhead[2] ((*well-formedness rules*+*additional_operations*)/*element_count*)) [7].

Given these less conclusive results (compared to model size analyses) it would be easy to be skeptical about the actual advantages offered by MLM. However, it is important to observe that these metrics were originally designed to target the type level only and thus entirely ignore the instance-level complexity caused by the application of TLM workarounds. This weakness of classic metrics for evaluating MLM is understandable given their motivation rooted in programming and/or modeling software. In these contexts, instances and their relationships are irrelevant to users. However, in many domain modeling applications instances directly represent the subject under study. In such contexts, the complexity of instance models is therefore very much a concern to users and should thus be considered in evaluations.

Instead of focusing on model properties (e.g., model complexity), one may also consider language properties (e.g., language expressiveness). For example, Atkinson et al. based their comparison of Melanee with MetaDepth on the differences between their respective language features [2]. Grossmann et al.'s more comprehensive comparison of 21 MLM approaches [9] also involved language fea-

[1] Apparently equivalent to the much debated "lines of code" metric for source code.

[2] Referred to as "complexity" in [7].

ture comparisons. However, Grossmann et al. also considered the intended target audience and the purpose of approaches, and furthermore considered the extent to which an approach has seen industry usage. This latter consideration could be regarded as including an external quality, but without further information on how well the respective MLM approaches performed in industrial contexts it is only a good starting point for further investigations.

Ideally, feature-based comparisons should be accompanied by an analysis of the impact of the different features on users. While certain features may seem elegant, ultimately their value must be assessed by considering external qualities.

5 Assessing External Qualities

In order to evaluate the ultimate purpose of any approach intended to deliver value to a user, it is necessary to determine properties based on external qualities which relate to user experience. As far as we are aware, only two MLM evaluations of this kind have been performed to date. Both of these investigate model changes and thus can be reasonably regarded as evaluating (aspects of) maintainability. In his Ph.D. thesis, Gerbig performed a comparative model change analysis by counting the number of primitive change operations needed to respond to certain requirements changes [7]. It turned out that a homogeneous treatment of all classification levels and Melanee's emendation service [1] reduce the effort needed to change the multi-level version of the model compared to the two-level, EMF-based version.

Kimura et al., also used a change-based approach to compare Melanee, MetaDepth and EMF, with a particular focus on extensibility [11]. These kinds of analyses exhibit ideal measurability, reproducibility, impartiality, and pragmatic relevance. However, whether context relevance is adequately addressed depends on how representative the chosen models and editing operations are.

Another external quality which lends itself relatively straightforwardly to measurement is *model robustness*, i.e., the resilience of a model to user error. Here the goal would be to assess the likelihood of introducing errors when creating/maintaining models. In particular, in the context of MLM to TLM comparisons, one would expect a two-level model to suffer from more accidentally introduced errors than a corresponding multi-level model. TLM would only provide the same safeguards against the introduction of model inconsistencies if all the well-formedness constraints implied by MLM are transposed into the equivalent TLM models. One would still, however, expect a higher rate of well-formedness violations, since it is most likely easier to make mistakes in a lower level two-level model, compared to a higher-level multi-level model.

The final external quality we can cover here is *productivity*, i.e., the speed by which users can develop or make changes to models. The underlying hypothesis of what could be referred to as *cognitive challenge*-based evaluations is that modeler performance is a function of the adequacy of the language/tool used. The higher the adequacy of the language/tool, the better the modeler should perform when facing standard tasks.

To this end, we propose a "5C"-approach, comprising the following cognitive challenges:

C$_1$: Comprehend Demonstrate understanding of a model.
C$_2$: Complete Read an incomplete model and correctly add missing parts.
C$_3$: Critique Read a defective model and identify all issues.
C$_4$: Correct Read a defective model and address all issues.
C$_5$: Create Create a model from scratch for a specified purpose.

Assessing the adequacy of an approach would be performed by measuring completion speeds for representative concrete tasks of the above five kinds. If languages/tools actually yield different levels of productivity, one should expect to see differences in the C$_1$-C$_5$ completion measurements. Ideally, subjects should be chosen in such a way that results transfer to the intended user base in order to achieve demographic relevance. Full context relevance will be very hard to achieve with this approach as it is typically not feasible to work with realistically sized models in such experiments.

6 Conclusion

The goal of this position paper has been to provide a discussion of the issues involved when aiming to perform *meaningful evaluations* while providing a broad overview of the MLM evaluations that have been conducted to date. The number of already existing MLM evaluations is encouraging and each of them represents a very useful step towards growing MLM as a discipline. However, our discussion has shown that the evaluations performed until now are overwhelmingly focused on internal rather than external qualities. Hence their pragmatic relevance – in the absence of the demonstration of a strong correlation between the internal qualities they asses with the external qualities that matter to users – is limited.

It is natural that the first evaluations performed in an emerging field are focused on internal qualities, as these are usually much easier to asses than external ones. However, we believe that for a) the benefits of MLM to become convincing enough to generate serious interest from industry, and b) comparative evaluations to become useful enough to maintain the cohesion and momentum the research community requires, more user-oriented evaluations focusing on external qualities will be needed.

An important initiative in this regard is the "Bicycle Challenge" proposed by the MULTI 2017 workshop as a common sample scenario, allowing various MLM approaches to be compared based on an example with practical relevance. Ideally, more such benchmarks will be designed in the future along with agreed upon usage scenarios, e.g., involving subsequent extensions, detecting and removing defects, etc.

It will remain a challenge to distinguish models and usage scenarios that have context relevance from those that do not, but any attempts to move MLM evaluations towards directly assessing external qualities or to strengthen the confidence in hitherto only assumed correlations between internal and external qualities will represent significant steps forward.

References

1. Atkinson, C., Gerbig, R., Kennel, B.: On-the-fly emendation of multi-level models. In: Proceedings of the 8th European Conference on Modelling Foundations and Applications. pp. 194–209. ECMFA'12, Springer (2012)
2. Atkinson, C., Gerbig, R., Lara, J.D., Guerra, E.: A feature-based comparison of melanee and metadepth. In: Proceedings of the 3rd Workshop on Multi-Level Modelling co-located with the 19th ACM/IEEE International Conference MODELS 2016. CEUR Workshop Proceedings, vol. Vol-1722, pp. 25–34
3. Atkinson, C., Kühne, T.: Reducing accidental complexity in domain models. Software and Systems Modeling 7(3), 345–359 (Springer Verlag, 2008)
4. Bansiya, J., Davis, C.G.: A hierarchical model for object-oriented design quality assessment. IEEE Trans. Softw. Eng. 28(1), 4–17 (Jan 2002), `http://dx.doi.org/10.1109/32.979986`
5. Chidamber, S.R., Kemerer, C.F.: A metrics suite for object oriented design. IEEE Transactions on Software Engineering 20(6), 476–493 (1994)
6. Coad, P.: Object-oriented patterns. Communications of the ACM 35(9), 152–159 (Sep 1992)
7. Gerbig, R.: Deep, Seamless, Multi-format, Multi-notation Definition and Use of Domain-specific Languages. Ph.D. thesis, University of Mannheim (2017)
8. Igamberdiev, M., Grossmann, G., Selway, M., Stumptner, M.: An integrated multi-level modeling approach for industrial-scale data interoperability. Software & Systems Modeling pp. 1–26 (2016)
9. Igamberdiev, M., Grossmann, G., Stumptner, M.: A feature-based categorization of multi-level modeling approaches and tools. In: Proceedings of the 3rd Workshop on Multi-Level Modelling co-located with the 19th ACM/IEEE International Conference MODELS 2016. CEUR Workshop Proceedings, vol. Vol-1722, pp. 45–55
10. Johnson, R., Woolf, B.: Type object. In: Martin, R.C., Riehle, D., Buschmann, F. (eds.) Pattern Languages of Program Design 3, pp. 47–65. Addison-Wesley (1997)
11. Kosaku Kimura, K.S.: An evaluation of multi-level modeling frameworks for extensible graphical editing tools. In: Proceedings of the 3rd Workshop on Multi-Level Modelling co-located with the 19th ACM/IEEE International Conference MODELS 2016. CEUR Workshop Proceedings, vol. Vol-1722, pp. 35–44
12. Lara, J.D., Guerra, E., Cuadrado, J.S.: When and how to use multilevel modelling. ACM Transactions on Software Engineering and Methodology 24(2), 12:1–12:46 (2014)
13. Lorenz, M., Kidd, J.: Object-oriented software metrics: a practical guide. Prentice-Hall, Inc. (1994)
14. Ma, H., Shao, W., Zhang, L., Ma, Z., Jiang, Y.: Applying oo metrics to assess uml meta-models. In: Baar, T., Strohmeier, A., Moreira, A., Mellor, S.J. (eds.) Proceedings of UML 2004, Lisbon, Portugal, pp. 12–26. Springer (2004)
15. Purao, S., Vaishnavi, V.: Product metrics for object-oriented systems. ACM Comput. Surv. 35(2), 191–221 (2003), `http://doi.acm.org/10.1145/857076.857090`
16. Rossini, A., de Lara, J., Guerra, E., Nikolov, N.: A comparison of two-level and multi-level modelling for cloud-based applications. In: Proceedings of ECMFA 2015. pp. 18–32. LNCS 9153 (2015)
17. Sommerville, I.: Software Engineering. Pearson, 10^{th} edn. (2016)
18. Yoder, J.W., Johnson, R.E.: The adaptive object-model architectural style. In: Proceedings of the 3rd IEEE/IFIP Conference on Software Architecture: System Design, Development and Maintenance. pp. 3–27. Kluwer (2002)