Quality criteria for mathematical models in relation to models' purposes

DOI:
10.1007/978-3-319-62968-1_12

Document status and date:
Published: 01/01/2017

Quality Criteria for Mathematical Models in Relation to Models' Purposes; Their Usefulness in Engineering Education

Jacob Perrenet
Industrial Engineering & Innovation Sciences, Eindhoven University of Technology
Bert Zwaneveld
Freudenthal Institute for Science and Mathematics Education, Utrecht University
Kees van Overveld (†)
Industrial Engineering & Innovation Sciences, Eindhoven University of Technology
Tijn Borghuis
Industrial Engineering & Innovation Sciences, Eindhoven University of Technology

Abstract:
A taxonomy of eight quality criteria for mathematical models was developed for the common basic modelling course in the innovated BSc curriculum of the Eindhoven University of Technology. First year engineering students of all disciplines reflected on their modelling group projects, answering the question how their models could be improved, using the criteria. The students were also asked to indicate the purpose(s) of their models from a list of 16 purposes. This study explores the purposes' and criteria's usefulness, defined as relevance combined with understandability. The purpose of *optimization* proved to be most relevant, followed by *analysis*, *prediction (what)*, and *verification.*The criteria of s*pecialization*, *genericity*, *scalability*, *distinctiveness*, and *convincingness* proved to be useful, the criteria of *audience, impact,* and *surprise* did not.

1 Introduction

For many years the Eindhoven University of Technology offered a series of about a dozen rather mono-disciplinary engineering bachelor programmes, such as Biomedical Engineering, Software Science, Applied Physics, and Applied Mathematics. Only in Applied Mathematics mathematical modelling had a prominent position (Perrenet and Adan 2011). In 2012 the totally innovated so-called Bachelor College started, comprizing all engineering Bachelor programmes as majors. On the one hand students were offered much more freedom to fill in their programmes, at the other hand a set of basic courses were obliged for all, including Modelling.

The aim of the modelling course is learning how to convert a non-mathematical problem into a form which can be tackled using mathematical tools, without losing sight of the original question. Starting with a generic introduction, the course offered four variants for specific disciplinary groups: dynamical systems (wherein time is the most important factor), data modelling (with acquired data as starting point), process modelling (concerning systems with distinct states), and modelling from scratch (not directly related to a specific discipline). The education period takes about 10 weeks with lectures, take-home assignments, intermediate tests, and a group project.

We will focus on the generic introduction of seven weeks. Its contents have been developed by one the authors of this chapter, the late Kees van Overveld, in consultation with another author, Tijn Borghuis, and with representatives of all the Eindhoven disciplines to make the content useful for all students. Many elements of the generic part were constructed especially for this course. In this explorative study we will zoom in on the topic of modelling criteria in relation to modelling purposes. We explore the usefulness of those criteria for the engineering students as apparent from their reflection on the modelling activities. Our aim is to share this part of Van Overveld's mental legacy with the mathematical modelling community. The development and implementation of the basic modelling course has been (and still is) a major operation with many content challenges (how to present modelling in a way that is acceptable for students as well as staff from many different disciplines?) as well as process challenges (how to teach over 1000 students, using lectures, lab sessions, and group projects?).

2 Theory

In ICTMA publications, respectively general mathematical education literature, much has been published about the *structure* of the mathematical modelling *process* and its quality, i.e., the steps to be performed and their order. Examples are Blum and Leiß (2006) and Girnat and Eichler (2011), respectively Blomhøj and Hoff Kjeldsen (2006) and Borromeo Ferri (2006). This structure is often represented by a modelling cycle; see Perrenet and Zwaneveld (2012) for an overview. There is an ongoing discussion about in how far these representations should be understood as descriptive or prescriptive. Perrenet and Zwaneveld (2012) conclude that as a criterion for the quality of the structure, most often is mentioned that validation as well as verification should be part of it.

Less attention has been given to the quality of the successful *result*: a working model, i.e., a model and its interpretation, with validation and verification taken for granted. Only a few lists of quality criteria for a working model can be found in literature. An example is the list of Meyer (1984) consisting of: accuracy (in how far is the model's output correct?), descriptive realism (in how far is the model based on assumptions which are correct?), precision (in how far are the model's predictions *definite* numbers, functions, geometrical figures, etc., or *a range of* numbers, etc.?), robustness (in how far is the model immune to errors in input data?), generality (in how far is the model applicable to a wide range of applications), and fruitfulness (in how far are the conclusions useful, inspiring, or pointing the way to another good model?). A second example is the list of Agoshkov (2002): adequacy (the extent of qualitative or quantitative agreement between the model and the modelled system concerning its properties), sufficient simplicity (balanced between giving reasonably accurate results that fulfill the stated purpose, and economy in terms of costs), completeness (yielding the best possibility, of obtaining desired outcomes), productivity (ease of measurement of input data), robustness (stability with respect to errors in input data), and clearness (direct, clear, substantial sense of a model's components).

Van Overveld and Borghuis developed a *taxonomy* of quality criteria that has some overlap with the lists mentioned before, but also takes into account a model's *purpose*.

A model's *purpose* also has not yet given much attention in modelling education research. (In this chapter we focus on higher education, Zwaneveld et al. (submitted) explore its use at secondary level). Van Overveld and Borghuis (2013) state that when we say 'one model performs better than another', 'better' should include 'with respect to the model's *purpose*'. They distinguish the following purposes (with typical questions added):

- *Explanation*: Why …? How comes …?
- *Prediction 1 (when?)*: When …?
- *Prediction 2 (what)*: What …? What if …?
- *Compression*: Can this data be summarized in fewer data or formulas?
- *Inspiration*: Maybe X could be tried…? Maybe Y could be true …?
- *Communication*: How to inform a specific audience??
- *Unification*: How to capture the similarity of phenomena from different domains?
- *Abstraction*: How to capture the essence of a phenomenon leaving out details?
- *Analysis*: What are the properties of the system under study?
- *Verification*: Is it true that …? (+ give argument)
- *Exploration*: What are the options …? In what ways can we connect A to B?
- *Decision*: Which of these is the best option (a closed set of alternatives)?
- *Optimization*: What is the best value for these parameters or dimensions?
- *Specification*: What external properties should some artefact have? What should it do?
- *Realization*: What internal properties should some artefact have? How should it do it?
- *Steering and control*: What (real time, online) interventions should this system do?

Three dimensions related to the modelling process produce an eight-criteria taxonomy:
a) the beginning (definition stage) or the end (conclusion stage);
b) the inside (model and modelled system) or the outside (stakeholders and context);
c) a qualitative or quantitative aspect of the model.

1. beginning/inside/qualitative: to which extent is the approach capable to handle various types of modelled systems or purposes? *Genericity*
2. beginning/inside/quantitative: to which extent can some characteristic dimensions of the problem increase, where the model still functions? *Scalability*
3. beginning/outside/qualitative: to which extent does the model / model outcomes requires specialized knowledge on behalf of the problem owner? *Specialization*
4. beginning/outside/quantitative: what size of intended audience does the model address? *Audience*
5. end/inside/qualitative: how plausible are the assumptions of the model? *Convincingness*
6. end/inside/quantitative: how similar can alternatives be in order for the model to allow distinction between these alternatives? *Distinctiveness*
7. end/outside/qualitative: What is the extent to which the model outcome may bring unforeseen new ideas? *Surprise*
8. end/outside/quantitative: What is the extent to which the model outcome can affect the stakeholders? *Impact*

Van Overveld and Borghuis stress that a complete taxonomy requires properties where all possible values can be enumerated. They develop operationalizations for all criteria. We will illustrate that here with two examples, impact and convincingness. *Impact* needs four quantities to express it as a number: r_1 = the profit or income in the present situation, without the model outcome; r_2 the profit or income with the model outcome in place; c_1 = the cost of ownership in the current situation; c_2 = the cost of ownership with the model outcome in place. For all quantities, the same time scale is taken (e.g., lifetime or yearly amounts). The quantity $\rho = ((r_2 - r_1) - (c_2 - c_1)) / (|r_2 - r_1| - |c_2 - c_1|)$ is a number between -1 and 1. Positive values mean a beneficial contribution; negative values mean that the impact is adverse. The absolute value $|\rho|$ indicates the size of the impact. *Convincingness* hinges on plausibility of assumptions. This can be related to an ordinal scale, from high to low, as follows. 5) Assumptions are logically deducible from other, non-problematic assumptions; 4) There is a law or theory in a well-accepted discipline (physics, economy, ...), such that the current assumption is an instance of that law or theory; 3) There is a plausible formal model system to which the current system may be compared; 2) There is an empirical model supporting the assumption and a similarity argument; 1) The model behaviour is consistent with intuition. See further Van Overveld and Borghuis (2013).

Van Overveld and Borghuis conclude with combining the purposes and the criteria. The relevance of the various criteria is related to the purpose(s) of a model. See Table 1.

Table 1 Relevance of criteria related to modelling purposes

criterion	genericity	scalability	specializ.	audience	convinc.	distinc.	surprise	impact
purpose								
explanation	x	-	x	x	x	x	x	x
prediction1	-	x	-	x	xx	xx	-	xx
prediction2	-	x	-	x	xx	xx	x	xx
compression	x	xx	-	xx	-	xx	-	x
inspiration	-	-	x	-	-	-	xx	-
communica.	xx	xx	xx	xx	xx	x	x	xx
unification	xx	-	-	-	xx	x	xx	-
abstraction	-	-	-	-	xx	xx	xx	-
analysis	xx	xx	xx	-	xx	xx	x	x
verification	x	xx	-	-	xx	xx	-	xx
exploration	xx	-	x	-	-	x	xx	
decision	x	x	x	-	xx	xx	-	xx
optimization	xx	xx	-	-	xx	x	-	xx
specification	xx	x	-	-	-	xx	-	x
realization	xx	x	-	-	-	xx	-	xx
steer.&cont.	-	x	x	x	x	xx	-	x

Note: communica. = communication; sterr.&cont. = steering&control; specializ. = specialization; convinc. = convincingness; distinc. = distinctiveness; xx = high relevance; x = medium relevance; - = no relevance.

To give one example: if the purpose is *communication*, the criterion *audience* (how large can the intended audience be?) is very relevant; if the purpose is *inspiration*, it is not.

Having elaborated the quality criteria for mathematical models in relation to their purposes, we can now formulate our explorative research questions:

- *In how far are the criteria useful for the students, i.e., relevant as well as understood?*
- *In how far are the purposes useful for the students, i.e., relevant as well as understood?*

3 Methods

3.1 Participants and materials

Our experimental subjects are 212 first and second year engineering student groups of various disciplines. The tasks used are embedded in regular education. In the beginning of their project reports the students had to state their modelling purpose(s). At the end they had to reflect as groups on their modelling project as follows.

Criteria reflection task 1) Necessity for improvement: In the lectures a set of criteria was presented to compare models. From the perspective of the model's purpose, on which of those criteria the model ought to be improved, according to your opinion and why? 2) Possibility for improvement: For which of those eight criteria do you have ideas about how to actually improve your model? Describe these ideas briefly.

3.2 Variables for analysis

We consider usefulness a two-sided concept, as well for the criteria as for the purposes. Firstly, a purpose or criterion should be relevant for the modelling problems the students meet. Secondly, the students should understand the concept. To account for both sides we define the *usefulness of a criterion* (or *purpose*) as the frequency of that criterion (or *purpose*) mentioned with understanding by the students minus the frequency of mentioned with misunderstanding. In both definitions frequency is the number out of all reports (212). A relative large positive score signifies usefulness; a low or even negative score signifies unusefulness (because of low use and/or because of low understanding). As a criterion we use that the name of a purpose or criterion is mentioned in relation to the context of the problem; from the description of the relation, understanding or misunderstanding is deduced; mixed understanding-misunderstanding is considered as misunderstanding. We will show some examples in the next section.

4 Results

4.1 Example problem with scoring of student groups answers

We present an example of a model assignment, completed with a scored example of student group answers to the purpose and criteria questions. Notation '------' is used for omitted text, ‘capitalized’ for use with understanding, and ‘underlined’ for use with misunderstanding.

> **Assignment: dynamic modelling project Virus infection**
> When a virus enters a human body it may replicate fast at first; our immune-system will react only after a threshold has been passed. If the body cannot cope with the virus-growth we need to administer an anti-viral drug, in due course, and in adequate amounts. Here is room for choices to be made. The possibilities also depend on the patient. Construct a model that indicates the results of a treatment as chosen by the specialist or general practitioner. Show how a responsible decision can best be made. Determine responsibility and what is best. Keep in mind that a model suitable for an adult may be inadequate for a child or infant.

> *Example of student group answers to assignment questions with scoring embedded:*
> “Purposes: The purpose of our model is to ANALYZE the disease. By analyzing how the virus behaves there can be control of the outcomes by using medication. To OPTIMIZE and CONTROL the amount of medication prescribed is also one of the purposes. The amount of medication can be controlled by following the advice our model offers. Optimizing the amount of medication leads to curing the infected and not spending too much money on medication. --------
> Possibilities for improvement: Our model could be improved for the categorizations: GENERICITY, <u>scalability</u>, ------. Our model can be improved in genericity because our model is now only for “normal” people. If more human aspects are taken into account, the model can be used for special cases as well. Our model does for instant not take in account that some people might be allergic to the medicine. ---- The scalability can be improved because we have no limit on the amount of medication. This is not realistic. There must be a state in which the person gets an overdose. We did not take that in account, so our scalability is too large for the model to be realistic in all situations. --------”

It should be noted, that we did not include the reference to a criterion’s operationalization, like described in section 3, as a requirement nor as a proof for understanding. We observed that students almost never used these operationalizations.

4.2 Validity and reliability

Scoring was mainly done by the first author. Before presenting the results, we will indicate the quality of this measurement.

Did we measure what we intended to measure? It should be noted that we only analysed limited sections of the project reports, i.e., those sections that should contain the answers to the specific questions to purpose and criteria according to the prescribed report format. If a question was not answered where it should be, the near context was scanned, but not the whole report. A small chance remains that the answers or other signs of specific understanding or misunderstanding were present elsewhere. As our indicators for usefulness are compound variables, scores sometimes might be ambiguous. If necessary however, this will be solved by split-up into understanding and misunderstanding.

Did we measure well what we measured? While the scoring of purposes posed few problems, the scoring of criteria was harder. In about 20% of the cases, the main rater had doubts; in those cases discussion followed with the other authors until consensus was reached. Especially because of this rather high percentage, drawing conclusions should be handled with care.

4.3 Results for Purposes

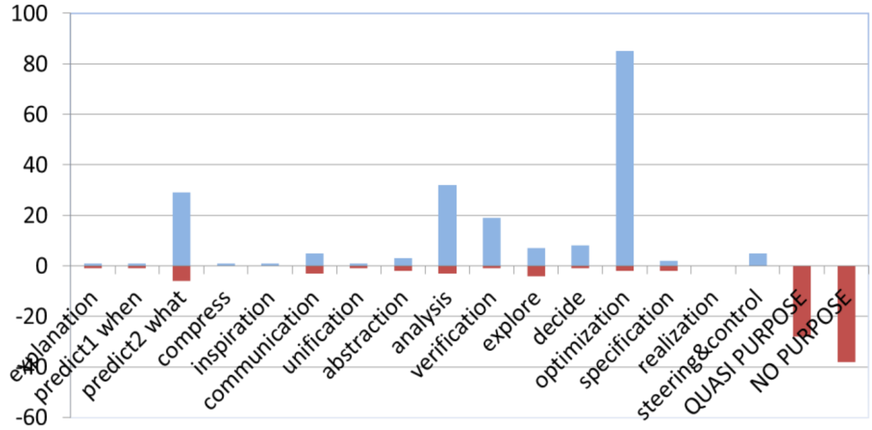


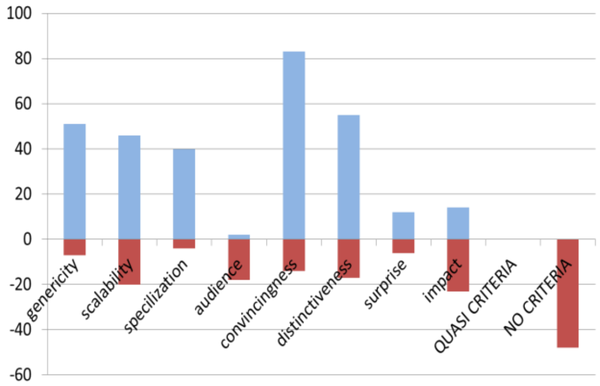
Fig. 1 Usefulness of purposes, split-up in understanding and misunderstanding

Figure 1 shows the frequency of use with understanding (out of 212) and under the axis the frequency of use with misunderstanding for all purposes. The data for quasi-use (other purposes) and non-use (no purposes) is enclosed. Clearly, the purpose of *optimization* is useful, followed by *analysis*, *prediction what*, and *verification*. All other purposes are hardly useful or not at all. Generally, when used, purposes are used with understanding. In other words, relevance is nearly equal to usefulness (total vertical length is nearly equal to length above the axis). Also the frequencies of cases where no purposes were mentioned or where other (quasi-) purposes were mentioned, are substantial. Examples of quasi-purposes are ‘determine’, ‘calculate’, and ‘investigate’. (We have to add at this point, that prediction was used often without the specification of ‘what’ or ‘2’, but because the other variant of prediction did not play any role, we did not see this as a sign for misunderstanding.)

4.4 Results for Criteria

Figure 2 shows the frequency of use with understanding as well as - under the axis - the frequency of use with misunderstanding (out of 212) for the eight criteria. The data for non-use (no criteria) and quasi-use (other criteria) are enclosed again. As we can see, the criterion of *convincingness* is clearly useful, followed by *genericity*, *distinctiveness*, *specialization*, and *scalability*. The other three criteria appear to be hardly useful (i.e., *surprise*) or even unuseful (*audience* and *impact*). Compared to the results concerning purposes, there is much more misunderstanding. *Audience* was often misunderstood as ‘being equal to *specialization*’, *impact* as meaning ‘influence’. Although *scalability* in general appeared to be useful, it showed some misunderstanding, such as putting it equal to *genericity* or giving it the meaning of ‘measurable on a scale’, and *distinctiveness* was sometimes given a meaning as ‘correctness’ or ‘as in reality’. Besides that, the frequency of cases where no criteria were mentioned appears to be substantial. Contrary to the case of purposes, no other (quasi-) criteria were mentioned. It should be noted here, that students sometimes mentioned such things as the group process that could or should be improved, but as it was not explicitly mentioned as a criterion, it was not considered a quasi-criterion.

Fig. 2 Usefulness of criteria, split-up in understanding and misunderstanding



5 Conclusions and Discussion

We will summarize our results:

– Only four of the 16 purposes are useful: especially *optimization* and also *analysis*, *prediction 2 (what)*, and *verification.*
– When purposes are mentioned, it is with understanding most of the time.
– With a substantial frequency other purposes than the formal ones are mentioned, such as determine, calculate, and investigate.
– With a substantial frequency no purpose is mentioned at all.
– Five of the criteria are useful: especially *convincingness* and also *distinctiveness*, *specialization*, and *scalability*. The other three criteria appear to be hardly useful (i.e., *surprise*) or even unuseful (*audience* and *impact*).
– When purposes are mentioned, it is sometimes with misunderstanding, such as audience as equal to specialization, scalability as equal to genericity, impact as meaning influence, scalability as meaning measurable on a scale, and distinctiveness as meaning correctness or as in reality.
– No other criteria than the formal ones are mentioned.
– With a substantial frequency no criterion is mentioned at all.

The result that there is more misunderstanding of criteria than of purposes might be explained by the fact that purposes are asked for in the beginning of the report, while criteria at the end, so possibly without tutor feedback. Also the criteria were new for staff.

Misunderstanding of criteria mainly falls into two categories: 1) blurring the quantitative and the qualitative aspect (e.g., *specialization* and *audience*); 2) blurring with familiar concepts having the same name (e.g., *impact* and *influence*).

As there were no specific instructions to staff members of the various departments creating problems, one might suppose that *optimization* and the other useful purposes were purposes that first came to mind when constructing a problem. Especially concerning *optimization* and also *analysis*, one might suppose they are in the kernel of engineering.

As the relevance of criteria is related to the purpose concerned, one might predict from Table 1 (see section 3), that the relevance of *audience* and *surprise* would be low and the relevance of *genericity*, *scalability*, *convincingness*, *distinctiveness* would be high, as they are. However, it would not predict high relevance for specialization and low relevance for impact. It might even be the case, that Table 1 (from Van Overveld and Borghuis 2013) could be improved, e.g., at first sight one could defend that *specialization* is relevant for *optimization* after all (to facilitate the client).

Our overall conclusion is that the criteria for comparing models related to the model's purpose as developed by Van Overveld and Borghuis (2013) are interesting and inspiring, but for optimal use in higher education some fine-tuning is necessary still: some terms are a bit ambiguous, some criteria overlap, and some relations between criteria and purposes might actually differ somewhat from as originally presented.

Acknowledgements

We thank the late Kees van Overveld for sharing the ideas treated in this chapter.

References

Agoshkov, V.I. (2002). *Mathematical Models of Life Support Systems*, Vol.1; UNESCO, Encyclopedia of Life Support Systems.

Blomhøj, M. & Hoff Kjeldsen, T. (2006), Teaching mathematical modeling through project work. *ZDM, The International Journal on Mathematics Education*, *38*(2), 163 – 177.

Blum, W. & Leiß, D. (2006). How do students and teachers deal with modeling problems? In C. Haines, P. Galbraith, W. Blum & S. Khan (Eds.). *Mathematical Modeling - ICTMA12: Education, Engineering and Economics* (pp. 222 – 231). Chichester: Horwood Publishing.

Borromeo Ferri, R. (2006). Theoretical and empirical differentiations of phases in the modeling process. *ZDM, The International Journal on Mathematics Education*, *38*(2), 86 – 95.

Girnat, B. & Eichler, A. (2011). Secondary Teachers' Beliefs on Modeling in Geometry and Stochastics. In G. Kaiser, W. Blum, R. Borromeo Ferri, & G. Stillman (Eds.), *Trends in teaching and Learning of Mathematical Modeling* - ICTMA 14 (pp. 75 – 84). New York: Springer.

Meyer, W.J. (1984). *Concepts of Mathematical Modeling*. New York: McGraw-Hill.

Overveld, Kees van, & Tijn Borghuis (2013). *From Problems to Numbers and Back; Lecture Notes to A Dicipline-neutral Introduction to Mathematical Modelling*; Internal publication Eindhoven University of Technology.[1]

Perrenet, J.C. & Adan I.J.B.F. (2011). Fifteen Years of Experience with Modelling Courses in the Eindhoven Program of Applied Mathematics; Journal of Mathematical Modelling and Application, *1*, 4, 3-19.

Perrenet, J. & Zwaneveld, B. (2012). The many faces of the mathematical modelling cycle. *Journal of Mathematical Modelling and Application*, *1*(6), 3–21.

Zwaneveld, B., Perrenet, J., Overveld, K. van & Borghuis, T. (submitted). *Mathematical Modelling in Dutch textbooks. Is it genuine mathematical modelling?*

[1] Obtainable by request to Tijn Borghuis (v.a.j.borghuis@tue.nl)