How to Assess the Quality of Student-generated Qualitative Models during an Open Modelling Task?

Marco Kragten[1[0000-0001-7194-3696]], Tessa Hoogma[1] and Bert Bredeweg[1,2[0000-0002-5281-2786]]

[1]Amsterdam University of Applied Sciences, Faculty of Education, Amsterdam, Netherlands
[2]University of Amsterdam, Faculty of Science, Informatics Institute, Amsterdam, Netherlands
{m.kragten,t.e.hoogma,b.bredeweg}@hva.nl

Abstract. Students often struggle with constructing models of system behaviour, particularly in open modelling tasks where there is no single correct answer. The challenge lies in providing effective support that helps students develop high-quality models while maintaining their autonomy in the modelling process. This study presents a procedure for assessing the quality of student-generated qualitative models in open modelling tasks, based on three characteristics: correctness, parsimony, and completeness. The procedure was developed and refined using student-generated models from two secondary school tasks on thermoregulation and sound properties. The findings contribute to the development of automated support systems that guide students through open modelling tasks by focusing on quality characteristics rather than adherence to a predefined norm model.

Keywords: Qualitative models, Quality assessment, Automated support.

1 Introduction

Frameworks for secondary education curricula emphasize the importance of students learning about system behaviour by modelling [1]. Through modelling, students not only develop domain-specific knowledge and modelling skills but also acquire epistemological understanding of the nature of models, their purposes, the modelling process, and the evaluation of models [2,3]. A common approach to modelling involves students using simulation software with a formal language to describe system behaviour [4-7]. However, students often struggle with formal expression of system behaviour [5,8], and limited domain knowledge leads to trial-and-error strategies rather than constructive modelling, resulting in minimal learning [9].

Qualitative models are promising for learning to model system behaviour. They describe system behaviour using a symbolic, non-numerical vocabulary that aligns closely with everyday human reasoning [10]. Previous research has shown that constructing qualitative models helps students understand domain-specific systems while simultaneously developing generic modelling skills [11]. Current modelling software typically offers automated support to students during model construction by comparing their models to a predefined norm model [12,13]. This support is triggered

when the models created by the students deviate from the norm model, which serves as the correct reference for the modelling task.

A crucial next step involves enabling students to construct their own qualitative models without a predefined norm model. In such a modelling task, "the right answer" is not fixed; students must determine how best to describe the domain-specific behaviour of the system using the formal vocabulary. This shift is essential for fostering deeper understanding of modelling and models, as students engage with the epistemological challenges of model construction [2,3]. However, supporting students in such open-ended tasks remains a challenge – especially as students' models may vary widely and evolve unpredictably.

This paper presents a procedure for assessing the quality of student-generated qualitative models, serving as a foundation for providing targeted support. Section 2 details the vocabulary specific to qualitative models in DynaLearn, the software used in this study. Section 3 discusses key quality characteristics and the types of knowledge required for validation. Section 4 describes the method, Section 5 the assessment procedure itself, and Section 6 the results. Section 7 presents the conclusions and discussion.

2 DynaLearn

DynaLearn (https://dynalearn.nl/) offers a qualitative vocabulary for modelling system behaviour (see Fig 1). Five distinct levels are designed to progressively support the modelling of increasingly complex behaviour [14]. This study focuses on level 2, a level commonly used in lower secondary education [15]. At this level, students work with five modelling ingredient types: entities, configurations, quantities, causal relationships, and values.

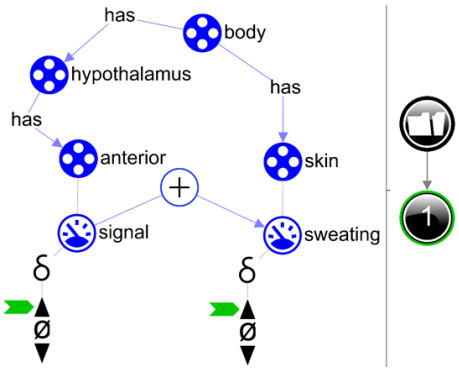

Fig. 1 Qualitative model describing a part of the thermoregulation system. The model (LHS) includes four entities: *body*, *hypothalamus*, *anterior*, and *skin*. These entities are structurally related through three configurations; for example, *hypothalamus* '*has' anterior*. The model also features two quantities: *signal* and *sweating*. There is a positive relationship between these quantities. The initial value of impulse is set to increasing (not shown).

The state graph (RHS) shows that the simulation results in a single state: as *signal* increases, sweating also *increases*.

Entities represent physical objects or abstract ideas within the system. They can be structurally related through **configurations**. Entities can have **quantities**, which are changeable properties that can increase, remain constant, or decrease. Quantities can have **causal relationships**. At level 2, these are categorized as positive (+) or negative (−) relationships. A positive proportional causal relationship signifies that the change in the affected quantity occurs in the same direction as the change in the causing

quantity. Conversely, a negative proportional causal relationship indicates that the changes are in opposite directions, e.g., if the causing quantity increases, the affected quantity decreases. The model can be simulated, using qualitative reasoning algorithms, which infer possible system states based on a given scenario [16]. Students must define a scenario by assigning **initial values** to the system. At level 2, this involves specifying the initial changes for quantities at the start of the causal chain.

3 Quality characteristics and knowledge sources

The quality of a student-generated model can be assessed based on three characteristics correctness, parsimony and completeness [17]. *Correctness* ensures that the model accurately represents the system being described. *Completeness* means that all necessary components are included so that the model sufficiently captures the intended system behaviour. *Parsimony* ensures that the model does not contain redundancies, making it as simple as possible while still being scientifically accurate. To assess these characteristics, we use three types of knowledge. *Domain knowledge* is required to assess whether a model accurately represents the system being described. This includes understanding which concepts belong to different ingredient types within the qualitative vocabulary. For instance, *sound* is an entity, while *amplitude* is a measurable property of sound and thus a quantity. *Knowledge of the vocabulary* is required to assess whether a model adheres to the formal rules of qualitative modelling, maintaining logical consistency. For example, all quantities in a model should be meaningfully integrated through causal relationships. *Knowledge about the purpose* of the model is necessary to assess whether the model includes appropriate details for the given learning context. A model that includes all possible relevant information may be scientifically correct but pedagogically ineffective if it overwhelms the student with excessive detail. For example, a thermoregulation model for lower secondary students should include heat loss and gain mechanisms but may omit molecular-level details.

4 Method

Participants. The study involved 24 lower secondary school students from two schools. At school A, 10 students worked on the sound task (see below for descriptions of the tasks), while at school B, 14 students worked on the thermoregulation task. They worked independently for 30 minutes, with access to instructional videos explaining model construction, a printed help sheet detailing ingredient functions, and the opportunity to ask questions. In a prior lesson, students had constructed a level 2 model on a different topic with norm-based support, ensuring familiarity with the modelling ingredients. Sessions took place on school premises but outside the regular classroom setting, with two students participating simultaneously. A research team member facilitated the sessions, providing dedicated support and maintaining a controlled environment free from classroom dynamics or peer distractions. The study was approved by the institutional ethics committee, and all participants and their legal guardians gave informed consent.

Modelling tasks. We designed two tasks, both topics commonly in lower secondary education. The *sound task* described how sound originates from a source (e.g., a guitar

string), travels through a medium (air), and is detected by a receiver (e.g., the human ear). Key properties included amplitude, which determines perceived volume and is influenced by the force of the guitar stroke, and frequency, which affects pitch and depends on the string's length, thickness, and tension. Students were required to model how variations in these properties impact the observed sound at the receiver. The *thermoregulation task* required students to construct a model of how the human body maintains internal temperature. The system description included temperature sensors in the skin that detect changes in temperature and the hypothalamus, which regulates body temperature by responding to these signals. The anterior hypothalamus triggers heat dissipation responses such as sweating and vasodilation, while the posterior hypothalamus induces heat conservation responses like shivering and vasoconstriction.

5 Assessment procedure

Figure 2 presents the general workflow for assessing the quality of student-generated models. The assessment process consists of two levels: ingredient assessment and whole-model assessment. At the ingredient level, when a student creates an ingredient, its correctness is assessed first, followed by the assessment of its parsimony and then its completeness. Once an ingredient has passed the assessment, the procedure moves to the whole-model level, where the completeness of the overall model is assessed.

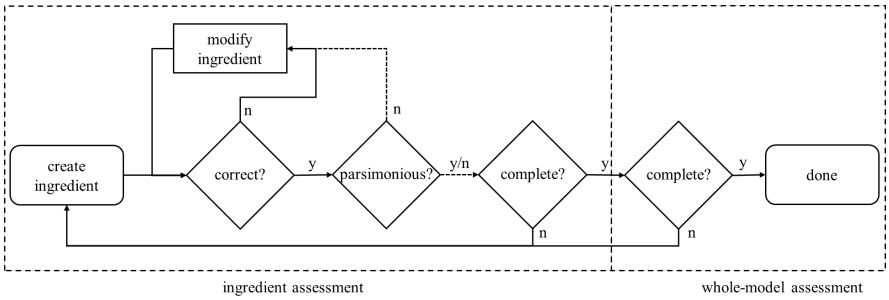

Fig. 2. Procedure for assessing the quality of student-generated qualitative models.

The specific checks required for each assessment step depend on the *type* of ingredient. Table 1 outlines these checks using the three quality characteristics: correctness, parsimony, and completeness. For example, *correctness* of **entities** is determined by two checks: (*i*) whether the entity has a name, as newly created ingredients should be given a name, and (*ii*) whether it is either a physical object or an abstract concept. The former check relies on knowledge of the vocabulary, while the latter requires domain knowledge. Note that, the checks are ordered, for instance, *name* assessment precedes *type* assessment. This helps managing the support, such as prompting students for 'giving a name' before asking about parsimony.

Next, *parsimony* is assessed. First, its relevance to the system is evaluated, requiring both domain knowledge and vocabulary knowledge. The entity should also represent a singular concept. For example, naming it *Ear is a receiver* combines two concepts, making it not singular. Additionally, uniqueness is evaluated to determine whether the entity appears more than once in the model. For instance, including two entities named

blood vessel may be unnecessary. However, unlike correctness, parsimony checks do not necessarily mandate changes. Finally, *completeness* checks whether the entity is related to another entity or has a quantity. This assessment relies on knowledge of the vocabulary. Similar checks hold for each model ingredient. See Table 1 for details.

Table 1. Quality checks for each ingredient type, associated knowledge type, and student-generated model scores for the sound thermoregulation tasks.

Ingredient	Quality	Check	Short description	Knowledge	S_t	T_t
Entity				*n* =	*21*	*89*
	Correct	Name	Has name	V	21	89
		Type	Is physical object/abstract idea	*D*	21	84
	Parsimony	Relevant	Is relevant	*D+P*	21	84
		Singular	Is singular concept	*D*	17	69
		Unique	Is unique	*D*	21	79
	Complete	Related	Has quantity/configuration	*V*	21	84
Configuration				*n* =	*7*	*54*
	Correct	Name	Has name	*V*	7	46
		Type	Describes structural relation	*D*	3	3
		Consensus	Scientific consensus	*D*	3	3
		Direction	Description in correct direction	*D*	2	1
	Parsimony	Unique	Is unique	*D*	2	1
Quantity				*n* =	*69*	*97*
	Correct	Name	Has name	*V*	69	92
		Type	Is measurable	*D*	57	57
		Entity	Is property of its entity	*D*	54	49
	Parsimony	Singular	Is singular concept	*D*	40	44
		Unique	Is unique	*D*	50	41
		Relevant	Is relevant	*D+P*	51	42
		Atomic	Quantities separated	*D+P*	53	47
	Complete	Related	Related to other quantity	*V*	48	38
Causality				*n* =	*56*	*62*
	Correct	Type	Both quantities of correct type	*D*	29	20
		Loop	*No* feedback loop	*V*	29	20
		Consensus	Scientific consensus	*D*	29	12
		Direct	Is direct effect	*D*	29	12
		Direction	Effect in correct direction	*D*	29	12
		Sign	Sign is correct (+/-)	*D*	29	10
	Parsimony	Relevant	Is relevant	*P*	29	10
Values				*n* =	*9*	*15*
	Correct	Conflict	Initial values *do not* conflict	*V*	9	14
	Parsimony	Redundant	*No* initial values in causal chain	*V*	9	13
	Complete	Initial	Initial value at start causal chain	*V*	3	8

Note. S_t = Sound task scores (N = 10); T_t = Thermoregulation task scores (N = 14); D = Domain knowledge; P = Knowledge of the purpose; V = Knowledge of the vocabulary; Scores of student-generated models for values are presented on model level (see text).

6 Results

We applied the assessment procedure to score the student-generated models for the sound task and the thermoregulation task. Table 1 presents the total scores. Figure 3 shows a student-constructed model. We use this to illustrate the assessment procedure. The model includes five **entities**: *body*, *blood*, *cold and heat regulation*, *hypothalamus*, and *sweat gland*. These entities correctly represent physical objects or abstract concepts and are relevant, singular, and unique. Each entity is connected, either through a configuration or an associated quantity. There are two **configurations**: one between *body* and *blood* and another between *body* and *cold and heat regulation*. However, these configurations lack assigned names, which prevents further assessment of their correctness. There are 6 **quantities.** The entity *cold and heat regulation* has temperature associated with it. While its type is correct (i.e., it is measurable), temperature is not a measurable property of this entity, making it incorrect. The entity *hypothalamus* has two quantities named *electrical signals*. While both are correct, singular, and relevant, one is redundant, making the model less parsimonious. Additionally, the quantity *electrical signals* on the upper right of the model is not related to any other quantity, making the ingredient incomplete. The quantity *sweat production* is correct, parsimonious and complete. The quantity *temperature* of the entity *sweat gland* is correct, but it is not relevant within the context of thermoregulation. In contrast, the quantity *temperature* of the entity *blood* is correct and relevant. There is a positive **causal relationship** between *temperature* of the entity *cold and heat regulation* and *electrical signals* of the entity *hypothalamus*. Correctness of this relationship cannot be assessed as the quantity *temperature* is not a measurable property of *cold and heat regulation*. A positive causal relationship exists between *electrical signals* of the entity *hypothalamus* and *sweat production* of the entity *sweat glands*. This is scientifically correct, relevant, direct, and correctly assigned in terms of direction and sign. The positive relationship between *sweat production* and *temperature* of the entity *sweat glands* is scientifically correct, although the effect is minor, making the relationship less relevant. Finally, the positive relationship between *temperature* of the entity *sweat glands* and *temperature* of the *blood* is scientifically incorrect (the effect is negligible). The quantity *temperature* of the entity *cold and heat regulation* is at the start of the causal chain and has an initial **value** assigned. No initial values are set within the causal chain. This is correct and prevents conflicts. Note, even though *temperature* is not a measurable property of the entity *cold and heat regulation*, the assessment of initial values remains possible, as it relies on knowledge of the qualitative vocabulary.

Whole-model. The assessment of whole-model completeness also reveals gaps in the model. For example, the model lacks separate entities for the anterior and posterior hypothalamus, as well as an entity for *blood vessels* with an associated quantity for its *diameter*. Consequently, the model also misses the causal relationships that demonstrate how *signals* from the *hypothalamus* affect the *diameter* of *blood vessels*.

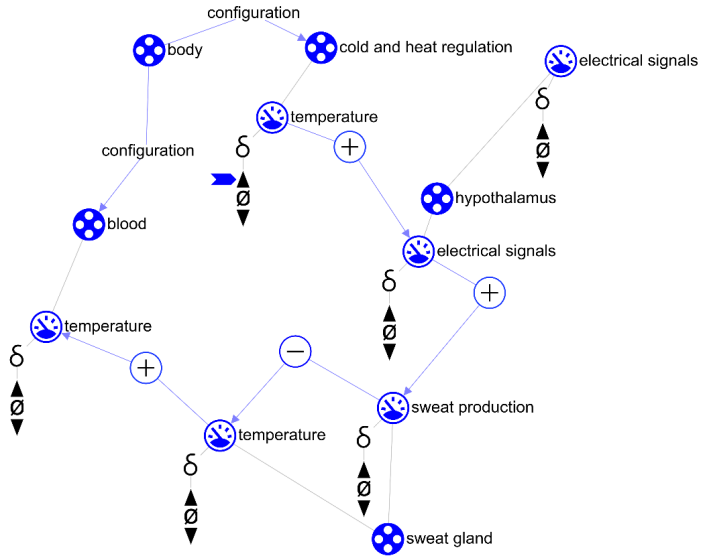

Fig. 3. Student-constructed model of the thermoregulation system.

7 Conclusion and discussion

We present a procedure for assessing the quality of student-generated qualitative models of system behaviour. The procedure offers a structured way to assess individual modelling ingredients (such as entities, quantities, and causal relationships) as well as the completeness of the model. Modelling ingredients are assessed using checks for correctness, parsimony, and completeness, with each check linked to a specific type of knowledge – domain knowledge, vocabulary knowledge, or knowledge of the purpose of the model. The procedure was applied to student-generated models from two modelling tasks: one on thermoregulation and one on properties of sound. The results demonstrate its usability and provide insights into common student difficulties when constructing models. These insights are valuable for understanding how students apply qualitative reasoning, where they tend to make systematic mistakes, and which aspects of the modelling process may benefit most from support. As such, this paper lays the groundwork for developing automated support systems that guide students in open modelling tasks.

While the assessment procedure developed in this study (in principle) employs a content-agnostic approach, there are several limitations to consider. The sample size was relatively small, which limits the generalizability of findings related to student performance and the specific modelling challenges observed. Although the procedure itself is not tied to any specific domain and can be applied across topics, the feasibility of fully automating the assessment remains an open question. In particular, the use of AI to supply the knowledge required for checks on correctness, parsimony, and completeness introduces several challenges – such as the reliability of AI-generated feedback (e.g., missing subtle domain-specific nuances), alignment with the qualitative

modelling vocabulary, and ensuring that the support remains pedagogically appropriate. However, rapid developments in this field offer promising avenues for future application and scalability. An alternative, more lightweight solution may lie in prompting students with reflective questions during modelling. This approach is easier to implement, as it leverages student reasoning and teacher facilitation rather than complex back-end automation and may already yield meaningful improvements in model quality and student thinking.

Future research should focus on developing and testing automated feedback mechanisms based on this assessment procedure. Note that, an ideal support system should also guide students through the broader modelling process – constructing, testing, simulating, reflecting, and refining [9]. Future work should explore how automation can facilitate this cycle. Additionally, integrating automated assessment with teacher guidance remains an important direction to ensure that human expertise complements machine-driven validation [18].

References

1. National Research Council.: Next generation science standards: For states, by states. (2013).
2. Bielik, T., Opitz, S. T., Novak, A. M.: Supporting students in building and using models: Development on the quality and complexity dimensions. Education Sciences **8**(3), 149 (2018).
3. Krell, M., Krüger, D.: Testing models: A key aspect to promote teaching activities related to models and modelling in biology lessons? Journal of Biological Education **50**(2), 160–173 (2016).
4. Bredeweg, B., Kragten, M., Holt, J., Kruit, P, van Eijck, T., Pijls, M., Bouwer, A., Sprinkhuizen, M., Jaspar, E., de Boer, M.: Learning with Interactive Knowledge Representations. Applied Sciences, 13(9), 5256 (2023).
5. VanLehn, K.: Model construction as a learning activity: A design space and review. Interactive Learning Environments **21**(4), 371–413 (2013).
6. van Joolingen, W., de Jong, T., Lazonder, A. W., Savelsbergh, E. R., Manlove, S. A.: Co-Lab: Research and development of an on-line learning environment for collaborative scientific discovery learning. Computers in Human Behavior **21**(4), 671–688 (2005).
7. VanLehn, K., Wetzel, J., Grover, S., Van De Sande, B.: Learning how to construct models of dynamic systems: An initial evaluation of the Dragoon intelligent tutoring system. IEEE Transactions on Learning Technologies **10**(2), 154–167 (2016).
8. Sins, P. H., Savelsbergh, E. R., van Joolingen, W. R.: The difficult process of scientific modelling: An analysis of novices' reasoning during computer-based modelling. International Journal of Science Education **27**(14), 1695–1721 (2005).
9. Mulder, Y. G., Lazonder, A. W., de Jong, T., Anjewierden, A., Bollen, L.: Validating and optimizing the effects of model progression in simulation-based inquiry learning. Journal of Science Education and Technology **21**, 722–729 (2012).
10. Forbus, K. D.: Qualitative representations: How people reason and learn about the continuous world. MIT Press, Cambridge (2019).
11. Kragten, M., Bredeweg, B. Calcium Regulation Assignment: Alternative Styles in Successfully Learning About Biological Mechanisms. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, I. I. Bittencourt (eds.), AIED 2024, LNAI 14829, pp. 220–234. Springer (2024).

12. Bredeweg, B., Kragten, M., Holt, J., Vaendel, D., Hanse, J., Bloemen, S. Stargazing live! inspiring with real data in a mobile planetarium and learning through conceptual modelling. In C. Frasson, P. Mylonas, & C. Troussas (Eds.), ITS 2023, LNCS 13891, pp. 257-269. Springer (2023).
13. Leelawong, K., Biswas, G.: Designing learning by teaching agents: The Betty's Brain system. International Journal of Artificial Intelligence in Education **18**(3), 181–208 (2008).
14. Bredeweg, B., Liem, J., Beek, W., Salles, P., Linnebank, F. Learning spaces as representational scaffolds for learning conceptual knowledge of system behaviour. In M. Wolpers, P. A. Kirschner, M. Scheffel, S. Lindstaedt, V. Dimitrova (Eds.), EC-TEL 2010, LNCS 6383, Springer (2010).
15. Spitz, L., Kragten, M., Bredeweg, B. (2021). Learning Domain Knowledge and Systems Thinking using Qualitative Representations in Secondary Education (grade 8-9). 34th International Workshop on Qualitative Reasoning, Montreal, Canada (2021).
16. Bredeweg, B., Linnebank, F., Bouwer, A., Liem, J. Garp3 — Workbench for qualitative modelling and simulation. Ecological Informatics, 4(5-6), 263–281 (2009).
17. Liem, J.: Supporting conceptual modelling of dynamic systems: A knowledge engineering perspective on qualitative reasoning. PhD thesis, University of Amsterdam, The Netherlands (2013).
18. Kragten, M., Hoogma, T. E., & Bredeweg, B. Integration of a Teacher Dashboard in a Hybrid Support Approach for Constructing Qualitative Representations. In R. Ferreira Mello, N. Rummel, I. Jivet, G. Pishtari, & J. A. Ruipérez Valiente (Eds.), EC-TEL 2024, LNCS 15159, pp. 208–221. Springer (2024).