The 14th International Congress on Mathematical Education
Shanghai, 12th –19th July, 2020

# VALIDATING A MODELLING COMPETENCIES ASSESSMENT

Jennifer A. Czocher, Sindura Kandasamy, Elizabeth Roan
Texas State University

*This development and validation study is part of a larger project focused on exploring development of mathematical modeling competencies among STEM undergraduates. We share a new assessment targeting modeling competency that is appropriate for undergraduates in advanced mathematics.*

The teaching and learning of mathematics at the tertiary level often benefits from a modeling approach. Typically, empirical studies suggest positive gains for students (e.g., in self-efficacy or robustness of mathematical knowledge) who are exposed to mathematical modeling (Czocher, 2017; Czocher, Melhuish, & Kandasamy, 2019). However, the field faces difficulty synthesizing results because the tools for assessment often serve only a local need. Indeed, in a survey of research literature, Frejd (2013) found that the majority of assessments were not grounded in theory but rather based on ad hoc constructions, personal experience, or small-scale studies of student work. We view this finding as indicating a clear need for a valid, reliable instrument capable of measuring gains associated with instructional interventions. However, theoretically and methodologically, assessing learning of mathematical modeling is uniquely difficult because the modeling process is itself idiosyncratic. It is difficult to formulate modes of assessment that can target instructional objectives when the target skills are not unidimensional. It is with these sensitivities to empirical and theoretical foundations of the genre, that we share efforts to develop an instrument targeting modeling skills of undergraduate STEM majors.

## CONCEPTUAL FRAMEWORK

This work is part of a broader project aimed at understanding how to leverage modeling and applications problems to help undergraduate STEM majors learn to define mathematical problems from nonmathematical situations. For this project, we adopt a view of mathematical modeling as a process of rendering a non-mathematical problem about a real-world phenomenon of interest as a well-posed mathematical problem to be solved. We focus on the cognitive activities that facilitate the process (Kaiser, 2017) and utilize the mathematical modeling cycle to operationalize these activities as skills or competencies (Blum & Leiß, 2007; Czocher, 2016). Challenges to students' development of modeling skills stem from the cognitive and metacognitive complexity of blending mathematical knowledge with real-world knowledge (Czocher, 2018; Fauconnier, 2006; Stillman & Galbraith, 1998). Students struggle to define mathematical problems from real-world situations because there can be an overwhelming number of considerations. Success requires *understanding* and *structuring*: specifying a problem and separating the relevant factors from ones that can be safely ignored. They struggle because the assumptions they introduce warrant mathematics that may not be accessible to them in the moment. Success requires *mathematizing* appropriate quantities, identifying relations among them, and representations to compose them. Models do need to be analyzed, but that is typically handled in mathematics classes. Finally, students struggle to *interpret* and *validate* their results, which involves checking that the model is representative of the situation and articulating its limitations. These activities are typically referred to as modeling competencies (Blum & Leiß, 2007). Because

competencies are interconnected, potential interventions and assessment should target the reasoning underscoring student decision-making. That is, an assessment of the extent to which an individual has developed a competency should acknowledge a student's justifications for her choices and that her choices may diverge from a normatively correct answer (Czocher, 2019).

**INSTRUMENT DEVELOPMENT**

Frejd (2013) observed that about one-third of assessments were written multiple-choice tests based on Haines, Crouch, & Davis (2000). Items were designed to target a single aspect of the modeling process (e.g., asking clarifying questions, identifying variables) and distractor responses were either irrelevant to the construction of a model or consider only the real-world constraints. The "best" answer choice considers both real-world constraints and relevant mathematics. Despite the promise of the instrument, critiques have been raised. First, the question set was tested on a small sample of students, so its properties are unknown. Second, there is some disagreement as to whether the parallel forms are indeed comparable. Third, research advances have enabled generation of distractor choices that align with students' tendencies. Fourth, psychometric models for item analysis have become more accessible.

Our approach to item construction adhered to the following constraints: (a) base problems were relevant and authentic, in the sense that they mimicked problems encountered in the students' studies or public discourse (b) phrasing of question stems and items appealed to multiple sources of student content knowledge (see Stillman, 2000) (c) question stems should target aspects of competencies via alignment to specific indicators of modeling activity (see Czocher, 2016), and (d) distractor drew on previous research studies of the kinds of factors and justifications that students exhibit when modeling.

We developed a pool of 120 multiple choice questions (MCQs) belonging to 9 real-world scenarios and some selections from the original Haines et al. (2000) items targeting the MMC competencies, except *working mathematically*. The real-world scenarios were drawn from research and educational materials (e.g., GAIMME report; textbooks; published research; faculty syllabi) that are appropriate to STEM undergraduates post Calc 2. We sought scenarios that treated issues prevalent in today's society, involved situations in the sciences where differential equations might be used, or were suggested by informal interviews with STEM professors. Mathematical content ranged from arithmetic to algebra, to calculus and to systems of ordinary differential equations. We then drafted MCQs from each scenario, striking a balance between information provided in the scenario set-up (so that the problems the MCQs addressed were situated) and readability (so that multiple MCQ stems could follow, cognitively, quickly from the set-up). A variety of question stems were used (e.g., select the most/best/least; indicate the choice consistent with the assumptions) and responses were developed to have a single "best" answer with four distractors at varying degrees of reasonability (e.g., reasonability for a *structuring* MCQ might address (un)helpful assumptions to make). One MCQ is in in Figure 1 (scenario set-ups omitted due to space constraints).

| **Mathematizing (selecting representation)** | |
|---|---|
| Which of the following models of human population growth is consistent with **all** of the following assumptions?<br>• The human birth rate is proportional to the population present,<br>• There are sufficient resources (e.g., space and ample food) for the population to thrive, given | a. $\frac{dP}{dt} = k_1 P - k_2 \frac{P(P-1)}{2}$<br>b. $\frac{dP}{dt} = k_1 P - k_2 \frac{P^2}{2}$<br>c. $\frac{\mathrm{dP}}{\mathrm{dt}} = k_1 P + k_2 \frac{P(P-1)}{2}$ |

- People die of old age and also prematurely, for example, due to malnutrition or inadequate medical supplies. Deaths also occur due to unnatural causes such as communicable disease and violent crimes.
- Deaths are proportional to the number of two-party interactions
- $k_1$ and $k_2$ are the proportionality constants for birth rate and death rate respectively

d. $\frac{dP}{dt} = k_1 P - k_2 \frac{P}{2}$

e. $\frac{dP}{dt} = k_1 P - k_2 \frac{(P-1)}{2}$

Figure 1: Sample item targeting an aspect of mathematizing

**Content and Construct Validity:** Two mathematicians who teach differential equations to STEM majors addressed content validity of the items. They examined readability of the questions, adequacy of the answer choices, and correctness of mathematics. Three mathematics educators with expertise in mathematical modeling evaluated readability of the questions, adequacy of the answer choices, and whether items were appropriately categorized in terms of the competency it was intended to target. We implemented revisions and suggestions, eliminating MCQs that failed to be correct or sensible or that duplicated other items.

**Round 1 Testing (Feasibility):** 59 items were sorted onto 3 forms (Red, Yellow, Blue) to balance competencies, and tested for feasibility with 14 STEM undergraduates enrolled in courses requiring differential equations or modeling at a large American university. At least 4 students responded to each MCQ (some participants skipped questions). Each MCQ was followed by a set of feedback questions: *Did you find anything confusing about the wording of this question or the wording of the possible answers? Why did you select the answer you did? Would you have chosen something different that was not an available choice?* Each scenario was followed by a set of feedback questions: *Please comment on the authenticity of [the scenario]. Do you find it believable in the real world? Was there any information you wish you knew about [the scenario] that would have helped you answer the questions? Was there any mathematics knowledge you wish you knew more about that would have helped you answer the questions*? We examined whether the students' responses were conceivably correct (justifiable) based on the reasoning provided. We modified the MCQs so that those justifiable reasons would either no longer apply or else used the reasons to enrich detractors. We also amended scenario set-ups to include additional relevant (but not necessarily relevant-to-model-construction) information where multiple individuals indicated that there was missing information. Where possible, we left ambiguity in the scenario intact, but removed ambiguity from the answer selections.

**Round 2 Testing (Difficulty):** 63 items, including 59 from Round 1, 3 from Haines & Crouch (2004), and8 from Haines & Crouch, et al. (2000), were sorted onto two forms to balance competencies and scenarios (Pink & Green). A total of 350 undergraduate STEM majors enrolled in courses requiring differential equations or modeling at a large American university signed up to test the items, from which 35 completed Pink and 43 completed green.The mean item difficulty revealed that most items (76%) were moderately difficult ($0.20 < p < 0.70$). Seven items were too easy ($p > 0.7$) and eliminated. Nine items performed worse than chance ($p < 0.20$) and were flagged for restructuring and one item was too difficult ($p = 0.03$) and was eliminated. We conducted a difficulty advantage analysis to assess whether those who studied differential equations had a sizable advantage over those who did not. Nine items suggested an advantage for students who took differential equations. None of these items were used in Round 3. Six items suggested an advantage for those who had not taken

differential equation. Four of these items were used in Round 3 since they were not too difficult for those who had not yet completed differential equations.

To analyze distractor efficiency, we calculated the proportion of students who selected each distractor. Of 253 distractors (62 items had 4 distractors and 1 item had 5), a majority were selected by at least 5% of respondents. In 17 of these items, distractors were selected more often than the keyed option. These items were flagged as items with potential to be discriminating items among students with varying abilities or as potentially requiring restructuring.

After restructuring items according to the option analysis, 30 items were selected for Round 3 on the following basis: (i) item difficulty should be in the $0.20<p<0.70$ range (ii) items should be sorted onto two forms with comparable total difficulty (iii) each form should contain the same number of items for each competency (iv) items where students who studied differential equations would have an outsized advantage over those who did not should be excluded. There were 28 items satisfying all of the criteria and so we selected an additional 2 items that satisfied criteria (ii), (iii), and (iv) but had a difficulty of p=.19. Each form balanced 4 items targeting *Mathematizing*, 4 items targeting *Validating*, 4 items targeting *Structuring*, 2 items targeting *Understanding*, and 1 item targeting *Interpreting*. Two items had $p < 0.20$ since applying all 4 of the above criteria was too constraining, however their p-values were close enough to $0.20$ to be worth retesting.

| Purple | | | Orange | | |
|---|---|---|---|---|---|
| Item Label | Competency | p | Item Name | Competency | p |
| Carrying Capacity 3 | Mathematizing | 0.67 | Decay 3 | Mathematizing | 0.42 |
| Carrying Capcity 6 | Validating | 0.45 | Decay 5 | Understanding | 0.35 |
| Decay 4b | Structuring | 0.37 | Decay 1 | Structuring | 0.20 |
| Haines & Crouch 11 | Understanding | 0.19 | Decay 2 | Validating | 0.38 |
| Disease 7 | Interpreting | 0.47 | Disease 6a | Understanding | 0.21 |
| Disease 9 | Validating | 0.41 | Disease 12 | Validating | 0.56 |
| Lagoon 4 | Structuring | 0.20 | Lagoon 2 | Structuring | 0.47 |
| Lagoon 5 | Mathematizing | 0.40 | Lagoon 8 | Validating | 0.30 |
| Lagoon 12 | Validating | 0.37 | Lagoon 9 | Mathematizing | 0.34 |
| Moth 5 | Validating | 0.30 | Moth 1 | Structuring | 0.19 |
| Population 5 | Mathematizing | 0.40 | Moth 12 | Mathematizing | 0.45 |
| Recycling 1 | Understanding | 0.53 | Population 3 | Mathematizing | 0.45 |
| Recycling 4 | Structuring | 0.37 | Population 4 | Validating | 0.34 |
| Recycling 5 | Mathematizing | 0.20 | Recycling 3b | Structuring | 0.33 |
| Recycling 14 | Structuring | 0.40 | Recycling 17 | Interpreting | 0.50 |

Table 2: Items tested in Round 3 with their p-values from Round 2

**Round 3 Testing (Difficulty, Distractor, & Discrimination):** The Orange and Purple forms were administered to a sample of secondary (25) and post-secondary (289) students participating in an international mathematical modeling competition focusing on the use of differential equations to solve real-world problems[1]. Of the 314 students responding to the forms, 115 responded to Purple form before the competition and 88 after while 111 responded to the Orange form prior to the competition and 87 responded after. Because the instrument is still in development, inferences about gains would be premature and therefore analysis of the pre/post performance is beyond the scope of this report.. Thus, we chose to collapse the pre and post responses to conduct item analysis. In total, there were 135 valid responses (students answered 66% of the MCQs) to the Purple form and 139 valid responses to the Orange form. Due to a survey platform error, Recycling 4 (structuring) and Population 4 (validating) were omitted from the difficulty and discrimination analyses.

The mean item difficulty for the Purple form was $0.359$ ($SD = 0.126$), with $0.177 < p < .0595$. The mean item difficulty for Orange was $0.369$ ($SD = 0.129$), with $0.147 < p < 0.580$. Across both forms, four items each were identified as too difficult ($p < .20$) or as having borderline ($p \approx 0.20$). We again conducted a difficulty advantage analysis. Of the 135 students who responded to the Purple form and of the 139 respondents to the Orange form, 121 and 128 had studied differential equations, respectively. The mean item difficulty for these groups were calculated across both forms and the item-wise differences between the groups were examined. Across both forms, there seemed to be notable advantages (differences in p-values $> .15$) for those having taken differential equations for 3 items (Lagoon 4, Moth 5, Moth 12). A chi-square statistic on the differences in p-values indicated that the difference was significant ($\chi^2 = .046$) only for Lagoon 4. On Recycling 1, there was a notable disadvantage for those not having taken differential equations and a chi squared statistic on the differences confirmed a borderline significance ($\chi^2 = .050$).

Our distractor analysis treats 56 distractors on each form. Common benchmarks for a distractor to function properly are (a) at least 5% of examinees should select each of an item's distractors and (b) the discrimination correlation should be negative. Across both forms, all distractors were selected by at least one person, but seven distractors attending seven distinct items were selected by fewer than 5% of the examinees. On the Purple and Orange forms, 16.07% (9/56) and 14.29% (8/56) of the discrimination correlation, respectively, were positive. Thus, 78.6% (Purple) and 80.4% (Orange) of the distractors were functional. All items had at least two functional distractors. On difficult items, a greater proportion of respondents selected distractors rather than the keyed option. The option analysis revealed four items for which a distractor was chosen much more often than the keyed answer (H&C 11, Decay 4b on Purple; Lagoon 9, Decay 5 on Orange). In each of these cases the distractor could *possibly* be seen as a justifiably correct answer. These four items were rekeyed to include the popular distractor as correct for the discrimination analysis. They have subsequently been restructured. After rescoring, the overall mean item difficulty of the purple form was $0.4042$ ($SD = .181,\ N = 135$), with $0.23 < p < 0.72$. The overall mean item difficulty for the orange from was $.4635$ ($SD = .2233, N = 139$), with $0.18 < p < 0.58$.

To conduct discrimination analysis, the point-biserial correlation was calculated for both forms under the revised scoring. The item-total point-biserial correlation ($r$PBIS) reflects the extent to which higher

[1] The SIMIODE Challenge Using Differential Equations Modeling, hosted by the Systemic Initiative for Modeling Investigations & Opportunities with Differential Equations https://www.simiode.org/scudem

ability students are more likely than lower ability students to select the keyed option. Thus, for a multiple-choice item to function effectively, the *r*PBIS must be positive. When the *r*PBIS is positive but small, it does not discriminate sufficiently among higher- and lower-scoring examinees to contribute to the overall quality of the assessment (DiBattista & Kurzawa, 2011). Only one item from the Purple form (Lagoon 9) was negatively correlated with the total score and all items from Orange form were positively correlated with the total score. All other items, from the purple from, had *r*PBIS$>$ $0.20$ while two items, from the orange from, had $0 <$ *r*PBIS$< 0.20$ and were flagged as low-discrimination items.

We calculated Chronbach's alpha as a measure of instrument reliability for the two forms. Under the original scoring for the Purple form, $\alpha = .430$ with highly difficult problems ($p < .20$), and $\alpha = 0.454$ without the highly difficult problems (for all items with $p > .20$). There was not a significant increase in $\alpha$ even after removing the highly difficult problems. Under the original scoring for the orange from, $\alpha = 0.556$ with highly difficult problems ($p < .20$), and $\alpha = 0.616$ without the highly difficult problems (for all items with $p > .20$). There was not a significant increase in $\alpha$ even after removing the highly difficult problems. Revised scoring led to increased α across both forms – for the purple form $\alpha = 0.556$ and for the orange form $\alpha = 0.625$.

**Performance of Existing Items from the Literature:** Because, to date, the items developed by Haines & Crouch, et al. are the only items targeting both differential equations and mathematical modeling, we provide an analysis of their performance with our samples. We tested 8 items found in papers (Haines et al., (2000); Haines & Crouch, 2004) listed in Table AAA.

| Item Label | Competency | p-value | % selecting key/distractor | DE Advantage |
|---|---|---|---|---|
| H&C1 | Mathematizing | 0.42 | 40.6 | 0.02 |
| H&C2 | Understanding | 0.84 | 84.4 | -0.10 |
| H&C3 | Interpreting | 0.53 | 51.6 | -0.27 |
| H&C5 | Understanding | 0.23 | 22.6 (key)/ 64.5 (distractor) | 0.33 |
| H&C7 | Understanding | 0.23 | 22.6 key/32.3 (distractor) | 0.20 |
| H&C11 | Understanding | 0.19 | 18.8 key/62.5 (distractor) | -0.08 |
| H&C13 | Structuring | 0.83 | 80.6 | 0.07 |
| H&C14 | Mathematizing | 0.71 | 68.8 | 0.10 |

Table 3: Difficulty and distractor analysis of Haines, et al. items

Of the 8 items, 3 items had $p > 0.70$, and were considered to be too easy. These items targeted structuring, mathematizing, and understanding competencies from the following scenarios respectively: aircraft evacuation, grocery store checkout, and display of street name signs. One item had $p = .19$ and was flagged as having near-chance levels of difficulty. While only 18.8% of the students selected the keyed option (option b) for this item, 62.5% of the students selected a distractor (option e) as the answer. The remaining 4 items had $0.20 < p < 0.70$. Of these four, two items had $p = .23$ but had a noticeable advantage ($> 0.15$) for those who had studied differential equations. At the same time, a third item gave a noticeable advantage to those who did not take differential equations. Of these 8 items, only one (H&C 11, size for stroller wheels) was tested again and had $p = 0.177$, again being flagged as too difficult (17.7% selecting the keyed option versus 54.6% selecting the same distractor as in Round 2). It is worthwhile to recall that the H&C items were developed to have a very tempting distractor (i.e., considering only mathematical or only real-world issues) for partial credit. It is possible that because the distractor had to do with the "smoothness of the ride as felt by the child"

posed in the question stem that it was more popular than the keyed option, which is problematic from an assessment perspective. It is also possible that the question was ambiguous since it is not clear to which attribute of the tire (e.g., radius or thickness) "size" is referring to.

## DISCUSSION

The utility of the MCQ instrument lies in its potential to measure gains in competencies, as constructed and as conceptually construed for STEM undergraduates who have taken or are enrolled in differential equations. There are some theoretical considerations that merit discussion. First, it is not clear that the items can be uniquely categorized as addressing a single modeling competency. This is because the competencies are largely presumed to work together, not in isolation. For example, model validation can rely on real-world data or on agreement with assumptions made earlier in the modeling process. This is a limitation of the instrument format. Second, and relatedly, the MCQ format means there is no additional observational information to help resolve, for example, whether a student "actually did mathematizing or validating" in response to an item. Thus, we recommend that the MCQ be used to assess levels of modeling competence, in general, rather than whether individuals have improved on specific competencies.

We are cautious, but optimistic, for interpreting the reliability and utility of the scale. The small sample sizes and low item numbers affect the estimate of Chronbach's alpha. Thus, Chronbach's alpha may provide a major underestimate of reliability. We are also cautious about interpreting correlations because the items provide dichotomous information. Further, we recognize that the construct "modeling competence" is not unidimensional because it draws from multiple scenarios (domains of real-world knowledge), multiple mathematical domains, and targets potentially distinct competencies. Future rounds of testing will pursue factor analyses, multi-dimensional testing theory, and move into Item Response Theory for documenting item-level characteristic to mitigate some of these methodological limitations. Finally, measuring gains in individual competencies adopts an atomistic approach and does not guarantee increase in an individual's modeling capacity holistically (see Blomhöj & Jensen, 2003). However, targeting competencies in interventions and assessment can support efforts to direct mathematics instruction toward ways of reasoning and justifying that are strongly connected to independent, autonomous modeling of complex situations. Due to the small number of items per competency, we relegate subscale analysis to future rounds of testing. Despite the limitations, the capability of the instrument to detect gains in modeling competence is hopeful because no individual's score was too high and there were some items where students scored below chance suggesting there is room for interventions to target the competencies and that the items will discriminate well based on ability.

## References

Blomhöj, M., & Jensen, T. H. (2003). Developing mathematical modelling competence: Conceptual clarification and educational planning. *Teaching Mathematics and Its Applications*, *22*(3), 123–140.

Blum, W., & Leiß, D. (2007). How do students and teachers deal with modelling problems. In C. Haines, P. Galbraith, W. Blum, & S. Khan (Eds.), *Mathematical modeling: Education, engineering, and economics* (pp. 222–231). Chichester: Horwood.

DiBattista, D., & Kurzawa, L.(2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scolarship of Teaching and Learning*, 2(2).

Czocher, J. A. (2016). Introducing Modeling Activity Diagrams as a Tool to Connect Mathematical Modeling

to Mathematical Thinking. *Mathematical Thinking and Learning*, *18*(2), 77–106.

Czocher, J. A. (2017). How can emphasizing mathematical modeling principles benefit students in a traditionally taught differential equations course? *Journal of Mathematical Behavior*, *45*, 78–94.

Czocher, J. A. (2018). How does validating activity contribute to the modeling process? *Educational Studies in Mathematics*. https://doi.org/10.15713/ins.mmj.3

Czocher, J. A. (2019). Precision, priorities and proxies in mathematical modeling. In G. A. Stillman & J. P. Brown (Eds.), *Lines of Inquiry in Mathematical Modelling Research in Education* (pp. 105–124). Cham, Switzerland: Springer Nature Switzerland.

Czocher, J. A., Melhuish, K., & Kandasamy, S. S. (2019). Building Mathematics Self-Efficacy of STEM Undergraduates' through Mathematical Modelling. *International Journal for Mathematics Education in Science & Technology*, 1–28. Retrieved from https://doi.org/10.1080/0020739X.2019.1634223

Fauconnier, G. (2006). Conceptual blending. In *The Encyclopedia of the Social and Behavioral Sciences* (Vol. 190, pp. 400–444).

Frejd, P. (2013). Modes of modelling assessment-a literature review. *Educational Studies in Mathematics*, *84*(3), 413–438.

Haines, C., Crouch, R., & Davis, J. (2000). Mathematical Modelling Skills: A Research Instrument. In *Mathematical Modelling*. Hertfordshire, England.

Haines, C., & Crouch, R. (2004). Applying mathematics: making multiple-choice questions work. *Teaching Mathematics and Its Applications: International Journal of the IMA 24*(2-3), 107-113.

Kaiser, G. (2017). The teaching and learning of mathematical modeling. In J. Cai (Ed.), *Compendium for Research in Mathematics Education* (pp. 267–291).

Stillman, G. A. (2000). Impact of prior knowledge of task context on approaches to applications tasks. *The Journal of Mathematical Behavior*, *19*(3), 333–361.

Stillman, G. A., & Galbraith, P. (1998). Applying mathematics with real world connections: Metacognitive characteristics of secondary students. *Educational Studies in Mathematics*, *36*, 157–195.