

Quantitative modelling biology undergraduate assessment

Robert Mayes^a, Kent Rittschof^b, Joseph Dauer^{ibc} and Bryon Gallant^d

^aMiddle Grades and Secondary Education, College of Education, Georgia Southern University, Statesboro, GA, USA; ^bCurriculum, Foundations, and Reading, Georgia Southern University, Statesboro, GA, USA; ^cSchool of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE, USA; ^dPsychology, Georgia Southern University, Statesboro, GA, USA

ABSTRACT

The Quantitative Modelling Biology Undergraduate Assessment (QM BUGS Version II) assesses undergraduate biology students' quantitative modelling abilities and confidence. The assessment is intended to be given in undergraduate biology courses where instructors are engaging students in quantitative modelling within biological contexts. The assessment consists of 36 questions: 25 multiple choice questions addressing four subcategories within quantitative modelling understanding (Quantitative Act, Quantitative Interpretation, Quantitative Modelling, and Meta-Modelling) and 11 Likert questions addressing student confidence about modelling in biology within the four subcategories. QM BUGS assessments were piloted in multiple undergraduate biology courses at both a research intensive university and regional university in fall 2017 (QM BUGS I) and spring 2018 (QM BUGS II). Here we present the development and theoretical framework for the assessment, focusing upon reliability and validity evidence with respect to measures of student understanding and student confidence following administration of the QM BUGS II.



ARTICLE HISTORY

Received 26 December 2018
Accepted 3 June 2019

KEYWORDS

Undergraduate biology; quantitative reasoning; modelling; confidence; Rasch analysis

A number of national reports call for an increased emphasis on modelling and quantitative literacy in science, technology, engineering, and mathematics (STEM) education (AAAS, 2011; AAMC and HHMI, 2014; COMAP & SIAM Garfunkel & Montgomery, 2016; NGSS Lead States, 2013). These calls to action seek to move STEM education towards authentic science practices where students construct, test, and revise models as they strive to understand natural phenomena (Magnani, Nersessian, & Pizzi, 2012; Windschitl, Thompson, & Braaten, 2008). There has also been an effort to cultivate authentic science practices in students and develop their model-based reasoning skills and meta-modelling abilities by engaging them in the modelling process (Papaevripidou, Constantinou, & Zacharia, 2007; Svoboda & Passmore, 2011). For example, students may be asked to generate a model, develop a hypothesis and prediction based on the model, make observations in the real world to test their hypothesis, and revise their model based on the results – all with an end goal of being able to generate a 'defensible explanation for the way the natural

CONTACT Robert Mayes  rmayes@georgiasouthern.edu  Middle Grades and Secondary Education, College of Education, Georgia Southern University, PO Box 8134, Statesboro, GA 30460, USA

world works' (Windschitl et al., 2008). In this way, modelling serves as a form of scientific inquiry and also permits students to explore the nature and purpose of models, thereby developing metacognitive modelling awareness that is essential to becoming a modeller (Papaevripidou & Zacharia, 2015).

In the field of biology, quantitative models have taken on a major role given the explosion of available experimental data from the study of complex global problems and the development of software and inexpensive hardware that permit data analysis and simulation (Li, Willer, Ding, Scheet, & Abecasis, 2010). Accordingly, there is a call to have undergraduate biology courses infused with quantitative modelling in an attempt to enhance students' abilities to understand biology concepts. Unfortunately, there has not been a corresponding development of validated assessments to determine whether these approaches are in fact helpful in increasing students' understanding of biology concepts (Aikens & Dolan, 2014). Consequently, most papers published on quantitative biology have described innovative teaching methods with limited inference. Without such assessments, it is unclear how instructors can determine the current quantitative modelling abilities of their students and the impact of any pedagogical changes to improve these abilities.

Modelling takes on many forms, including experiential (physical manipulatives), visual, verbal (qualitative discourse), numerical (quantitative data), or symbolic quantitative models (Eaton et al., 2017). Multiple model representations can provide different views of a biological problem and thus have the potential to improve students' comprehensive understanding (Ainsworth, 1999; Stieff, 2017) as they encode and retrieve knowledge in different modalities (Paivio, 1990). The QM BUGS II assessment development was initiated as part of the Quantitative Undergraduate Biology Education and Synthesis (QUBES) Project. QUBES addresses challenges in quantitative biology education, providing support for innovative biology education (QUBES, 2019). One focus of the QUBES Project is quantitative models, which have recently become prominent in biology (Li et al., 2010). Quantitative models use mathematical concepts and language to describe phenomena. They include many forms, such as statistical models, function models, differential equations, game theoretic models, and logical models.

Computational modelling and simulation modelling are two examples of modern quantitative modelling approaches that permit a deeper understanding of underlying mechanisms and the ability to investigate a complex biological process as a whole (Robeva, 2015). Computational modelling is the use of computers to simulate and study the behaviour of complex systems using mathematics, physics, and computer science (National Institute of Biomedical Imaging and Bioengineering, 2018). Simulation modelling is the process of creating and analysing a digital prototype of a physical model to predict its performance in the real world (Winsberg, 2003). For example, quantitative modelling requires learners to develop a quantitative account of the phenomena and understand mathematical and conceptual interactions among the model components (Mayes et al., 2013). Translation across multiple representations is strengthened through quantitative interpretation of models when determining trends and making predictions (Mayes, Forrester, Christus, Peterson, & Walker, 2014). Importantly, students who are given the opportunity to develop the quantitative models themselves become owners of the modelling process since they are responsible for learning about the phenomena (Papaevripidou & Zacharia, 2015; Schwarz et al., 2009).

To address this knowledge gap, our long-term goal is to build a foundational knowledge base of research so all students have access to high-quality, inspiring quantitative biology teaching. Towards that end, the goal of the **Quantitative Modelling by Biology Undergraduate Students** (QM BUGS) project is to determine undergraduate students' quantitative modelling abilities in biology and the impact of modelling on their understanding of biological concepts. In support of this goal, our objectives are to:

- (1) Determine the relationships among quantitative modelling ability, modelling metacognition, and disciplinary knowledge;
- (2) Examine the impact of pedagogical modelling intensiveness on the development of students' quantitative modelling abilities;
- (3) Develop a validated diagnostic assessment to measure undergraduate students' quantitative modelling abilities and confidence in biology.

The focus in this paper will be a discussion of assessment development related to goal 3. The research team – comprised of a biology education researcher, mathematics education researcher, and educational psychologist – have developed a biology quantitative modelling diagnostic assessment, thereby providing a tool to determine the current state of development of students' abilities to create and reason with quantitative biology models.

Materials and methods

Assessment

The Quantitative Modelling Biology Undergraduate Student Assessment version II (QM BUGS Version II) assesses undergraduate students' quantitative modelling abilities and confidence in biology. The assessment is intended to be administered within undergraduate biology courses where development of quantitative skills is preparing students to actively engage in quantitative modelling within biological contexts. The assessment consists of 36 questions: 25 multiple choice questions addressing four subcategories within quantitative modelling understanding and 11 Likert questions on a 4 level scale (from 2 Strongly Disagree to 5 Strongly Agree) addressing student confidence about modelling in biology. A rating of 1, Not Applicable, was included as an opt-out but not considered part of the ordinal rating of confidence. The four subcategories within quantitative modelling understanding are Quantitative Act (QA Q1–Q6), Quantitative Modelling (QM Q7–Q13), Quantitative Interpretation (QI Q14–19), and Meta-Modelling (MM Q20–Q25). The confidence items include one QA confidence question (Q26), 7 QM confidence questions (Q27–Q33), and 3 QI confidence questions (Q34–Q36). The focus of these items is on confidence in modelling, but QA is foundational to engaging in modelling so one item is included, and full comprehension of a model requires QI so three items are included.

Implementation

The assessment has been through multiple development cycles. We provide a brief summary of the pilot assessments leading up to the development of the QM BUGS II assessment

which will be the focus of our discussion. In fall 2016 two separate quantitative modelling assessments were developed: QUBES QA-QL and QUBES QI-QM.

QM BUGS QA-QL assessment

The QA-QL assessment (fall 2016) focused on Quantification Act (QA – ability to move from biological context to quantitative account) and Quantitative Literacy (QL – ability to use simple arithmetic and algebraic methods to quantify relationships within a context, that is, to combine, compare, contrast, and manipulate the variables quantified). The assessment was designed to be given at the beginning of a course as a diagnostic assessment indicating areas where students need QA-QL just-in-time instruction or supplemental instruction, ensuring the student possess prerequisite skills for interpreting (QI) and building quantitative models (QM). The assessment consisted of 26 multiple choice questions in the following subcategories: Quantitative Act (Q1), QL Numeracy and Number Sense (Q2–Q5), QL Measurement (Q6–Q7), QL Proportional Reasoning (Q8–Q14), QA Probability and Statistics (Q15–Q26). Students were randomly assigned 10 items with at least one coming from each of the five subcategories. The purpose of the pilot was primarily to evaluate the QUBES QA-QL assessments. The decision was made to pursue quantitative modelling and no further development of a separate QA-QL assessment was conducted. However, QA items were incorporated into the QM BUGS II assessment, since QA is foundational to model development.

QM BUGS QI-QM assessment

The QI-QM assessment (fall 2016) measured students' abilities to engage in Quantitative Interpretation of biology models (QI – ability to interpret a model provided to them, for example, STEM literate citizens interpreting a science model to make informed decisions about an issue) and to determine students' proficiency and confidence in Quantitative Modelling (QM – ability to develop a model). The assessment was designed to be administered within undergraduate biology courses. The assessment consisted of 13 questions (6 multiple choice and 7 Likert scale items) in the following subcategories: QA Variable and Variation (Q1), QI (Q2–Q5), QM (Q6), and QI-QM Confidence (Q7–Q13). The pilot assessment results were used in determining major revisions of items.

QM BUGS I assessment

The QM BUGS I assessment (fall 2017) consisted of 26 items, 19 of which were multiple choice understanding questions focused on a broad set of guiding frameworks including Quantitative Reasoning Learning Progression (Mayes et al. 2014), Model-based Reasoning (Schwarz et al., 2008), and metacognition of modelling (Papaevripidou & Zacharia, 2015). The assessment consisted of questions in the following subcategories: Model Formulation (Q1–Q2, Q5–Q12), Model Deployment (application of model for prediction) (Q13, Q15–Q19), Modelling Reasoning (Q3–Q4, Q14), and QA-QI-QM Confidence (Q20–26).

QM BUGS I pilot. The QM BUGS I assessment was piloted in fall 2017 at the University of Nebraska-Lincoln in one section of NRES/BIOS 220 Ecology and Georgia Southern University in two sections of BIOL 3133 Evolution and Ecology and BIOL 4635 Animal Behaviour. The assessment was taken by students online through Qualtrics and was offered as an extra credit assignment for the classes. The assessments were completed in the final

two weeks of the semester for all the classes. There were 171 students that accessed the assessment. Of those, 20 students did not complete the assessment and 2 took the assessment twice. Those who did not complete the assessment were removed from the sample and only the final attempt for the 2 repeating the assessment were accepted, leaving 149 cases that were analysed.

QM BUGS I results. An initial item analysis for the 19 QM understanding items indicated that overall the assessment was relatively difficult for the students. Ten of the 19 items had the correct response as the most selected, but the percentages for selection were as low as 30.5% and high as 62.3%, with an average of 50.7%. The remaining 8 items all had an incorrect response selected more often than the correct response, often with a percentage difference greater than 20% favouring the incorrect item. Three of the 8 items (Q7, Q13, Q16) had more than one incorrect response chosen more often than the correct item, making them especially suspect. An analysis of the assessment items indicating potential issues was conducted to inform revision of the assessment. This included a Rasch analysis of the assessment. A detailed discussion of Rasch analysis is provided in the QM BUGS II review below. Evidence indicated that the low scores could have resulted from poor student understanding of quantitative modelling in biology, but that scores were also likely impacted by issues with the items. The QM confidence items focus on students' beliefs about their ability in modelling. Rasch Analysis did not indicate any of these items required revision. Extensive revision of the QM BUGS I assessment was completed for a second data collection in spring 2018.

QM BUGS II assessment

A number of theoretical frameworks on modelling influenced the development of the diagnostic assessments. The frameworks selected for QM BUGS II provided elements of modelling which guided assessment item development. The crosswalk of framework elements and assessment items is provided in Appendix A. The following were key frameworks driving the development of the QM BUGS II assessment.

- (1) The Quantitative Reasoning Learning Progression (Mayes et al. 2014) provides three progress variables each with four achievement levels defining characteristics of the variables. The progress variables are Quantification Act (QA) which includes quantifying the variables, situative view of context, covariation, and quantitative literacy; Quantitative Interpretation (QI) of models to determine trends, make predictions, translate between models, and revising models; and Quantitative Modelling (QM) including creating models, refining models to address new situations, reasoning with models, and statistical analysis.
- (2) The QUBES Modelling Framework (Dahlquist et al., 2018) provides a flow diagram for quantitative biology modelling identifying movement between the science experimental model design and the mathematical model design. Both designs begin with formulating a problem, pass through parallel conceptualization stages with science conducting experiments and mathematics building quantitative models, move to validation through science experiments or implementing mathematical models, then both conclude with analyse, interpretation within context, and dissemination of findings.

- (3) Model-based Reasoning (Krajcik & Merritt, 2012; Lehrer & Schauble, 2000; Louca & Zacharias, 2012 Schwarz et al., 2009;) focuses on the ability of students to construct models in order to explain observed phenomena. Conceptual elements include modelling based on observations, pattern identification, models as ideas not physical objects, acceptability and uniqueness of models, empirical or theoretical objects that constitute models, empirical and conceptual assessment of models, and models guiding future work.
- (4) The MoDeLs Project (Schwarz et al., 2008) provides an instructional modelling sequence consisting of presenting an anchoring phenomena, constructing a model of phenomena, empirically testing the model, evaluating the model, testing the model against other theories, revising the model, and using the model to predict or explain other phenomena.
- (5) Modelling Framework of Learning (Lehrer & Schauble, 2005; Louca & Zacharias, 2012; Metcalf, Krajcik, & Soloway, 2000; Nicolaou & Constantinou, 2014; Sins, Savelsbergh, & van Joolingen, 2005; Windschitl et al., 2008) provides a foundation for modelling-based learning (MbL). MbL is an approach for teaching and learning in science where students construct models as representations of physical phenomena. The models include representations of objects characteristics and processes to increase student understanding of the phenomena (Louca & Zacharias, 2012). Modelling practices including model formulation (Duschl, Schweingruber, & Shouse, 2007), model comparison (Pluta, Chinn, & Duncan, 2011), and model evaluation (Schwarz & White 2005; Snir, Smith, & Raz, 2003).
- (6) Metacognition of modelling includes thinking about the process of modelling, self-regulation through explicit identification, and description of major steps in modelling process (Papaevripidou & Zacharia, 2015). Major steps in the modelling process include model formulation through analysis of phenomena, inductive reasoning to hypothesize how variables interact, and quantifying to formulate a model; model deployment including documenting and empirically testing the model, evaluating the model, testing against other models; and meta-modelling including the nature and purpose of models, as well as steps of modelling.
- (7) Meta-modelling provides elements of the nature and purpose of models (Oh & Oh, 2010; Schwarz & White 2005). Meta-modelling includes self-regulation in identifying and describing major steps of modelling process (Papaevripidou & Zacharia, 2015), knowledge corresponding to understanding the nature of models (Schwarz & White, 2005), and appreciation of the purpose and utility of models (Oh & Oh, 2010).

Sample items from QM BUGS II are provided in Appendix B. One item was selected from each of the five sections of the assessment:

- Quantification Act (QA): Item 5 – quantitative literacy, building expressions
- Quantitative Interpretation (QI): Item 8 – create model, phenomenological
- Quantitative Modelling (QM): Item 18 – model application, prediction
- Meta-Modelling (MM): Item 24 – purpose and utility of models
- QM Confidence (QC): Item 26 – QA, Item 30 – QM, and Item 35 – QI

Table 1. Sample demographics.

	Group	Frequency	Per cent
School/Class	UNL	43	65.2
	GSU	23	34.8
Gender	Female	37	56.1
	Male	28	42.4
	Not Identified	1	1.5
Race	Black/African American	10	15.2
	Hispanic/Latin	6	9.1
	Caucasian	45	68.2
	Asian/Pacific Islander	4	6.1
	Native American	0	0
	Other	1	1.5
Grade	Freshman	23	34.8
	Sophomore	12	18.2
	Junior	9	13.6
	Senior	20	30.3
	Graduate	2	3.0

QM BUGS II implementation. The QM BUGS II assessment administration was conducted in spring 2018 at the University of Nebraska-Lincoln in LIFE 121 Fundamentals of Biology and at Georgia Southern University in Biology 5540: Ecology and Biology 1155 Comparative Animal Physiology. The assessment was taken by students online through Qualtrics and was offered as an extra credit assignment for the classes. The assessments were completed in the final two weeks of the semester for all the classes. QM BUGS II was administered to determine the current status of students modelling abilities and confidence as impacted by their biology programme. There was no adjustment required of instructors in the amount or approach to teaching quantitative modelling.

Demographics of the sample completing the QM BUGS II assessment are provided in Table 1. There were 80 students that accessed the assessment on the internet using a web browser. Of those, 13 students did not complete the assessment and one took the assessment twice. Those who did not complete the assessment were removed from the sample and only the first attempt for the student repeating the assessment was accepted, leaving 66 assessments that were analysed.

Results

QM BUGS II included 25 multiple choice questions addressing QM understanding. The correct response was the most often selected on 17 of 25 questions, with 11 of these selected over 56.1% of the time. Of the remaining 8 of 25 items, the correct response was the second most selected on 5 of these 8 items. The most problematic items were those where the correct response was chosen less often than two or more other distractor responses, which included Q11, Q14, and Q22. The overall group average on QM BUGS II represented a substantial improvement over that of QM BUGS I. This may be an indication of improvements in the clarity of the assessment and perhaps the instruction provided.

The 11 Likert confidence questions of QM BUGS II were analysed by examining the distribution of the response across the 4 level scale. A higher score corresponds to more reported confidence. Responses on the 11 items were relatively high with a mode of 3 on the

4 level scale, indicating a high level of confidence. There was no indication of problematic confidence items.

Rasch analysis of QM BUGS II

Rasch measurement methods were used to analyse both the student outcome measures and assessment item measures simultaneously (Bond & Fox, 2015; De Ayala, 2011; Engelhard, 2013; Linacre, 2012), permitting a close examination of the validity and reliability evidence relative to the QM BUGS II assessment process. A Rasch approach was chosen for these purposes because it allows the construction of additive measures from our data as we examine both item statistics and individual student statistics that inform revision of the assessment (Wilson, 2009). Rasch approaches are essentially a family of modern latent trait models, whereby each member model corresponds with a particular type of data. Rasch models include, among other models, the dichotomous model (Rasch, 1960/1980) for correct/incorrect data, and the rating scale model (Andrich, 1978) for Likert rating data, which were each used within this investigation. Both Rasch models used here provide an idealized basis for comparison of item statistics and person statistics constructed on the same uni-dimensional, linear scale. Rasch measurement allows us to take advantage of an item difficulty parameter and diagnostic statistics (Linacre, 2012) for evaluation of calibrated person measures, item measures, and the overall measurement scale in comparison with an idealized measurement model of that data. The Winsteps (Linacre, 2014) computer program was used for Rasch measurement calibration and the SPSS computer program was used together with Winsteps for subsequent diagnostic and comparative analyses.

Rasch analysis of QM understanding items

The Rasch dichotomous model (Rasch, 1960/1980) was used in order to construct linear measures from the correct/incorrect scoring used with the QM BUGS II assessment for understanding items (Q1–Q25), then to examine and improve measurement accuracy.

A primary Rasch calibration was run on QM BUGS II understanding items (Q1–Q25) to identify potential problematic items and students. The research team decided not to eliminate items from calibrations, so all 25 items were included in all analyses. Student responses were analysed to determine if displayed answer patterns indicated lack of fidelity in responding to items (for example long strings of the same response on assessment).

Person measures. Comparing raw data or percentages can be problematic when using multiple choice scales because it incorrectly assumes that each correct response is an equivalent point toward the total, leaving a misleading suggestion that a total or percentage score is equivalent to a measure. The assumption of equivalency is misleading because items have different levels of difficulty, making them non-equivalent. Rasch methods calculate a scale score from the raw scores, referred to as the ‘measure’, which address this problem. The measure uses logits (logarithm of the odds of agreement) as the unit with a mean of 0, and a typical range of approximately -3.0 to 3.0 logits. The measure is a standardized scaled score (Table 2) that is plotted on the Rasch ruler. The mean person measure of -0.11 indicates that students’ mean was slightly below the item difficulty mean for this analysis. The mean model error of 0.47 indicates a confidence interval of $(-0.58, 0.36)$ about the mean.

Table 2. Person summary for the QM BUGS II understanding items.

Person summary	QM BUGS II understanding items (Q1–Q25)	
	Total raw score	Measure
Mean	12.1	−0.11
Standard deviation	3.9	0.84
Max	20	1.68
Min	3	−2.37
Reliability	.67	0.65
Separation	N/A	1.37

Item polarity was examined to identify whether items function in unison as reflected by positive point measure correlations. Point measure correlations were all positive and most were relatively strong, that is above criteria of 0.50 and not less than 0.15. The exceptions were Q14 (correlation of 0.01) and Q19 (correlation of 0.05). Q9 (correlation of 0.27) and Q22 (correlation of 0.16) had low point measure correlations, but above the criteria used.

Reliability and validity indices. Rasch calibration calculates a person reliability index which is similar to Cronbach's alpha, except that it uses constructed measures rather than the raw scores used for alpha (Table 2), allowing it to be more accurate. The QM BUGS II assessment yielded person reliability levels of 0.65 indicating the assessment showed only a moderate level of reliability that was below the criteria of .80 sought. Separation indices of 1.37 reflects weak distinction in levels of students. The Total Score columns provide a student raw score summary, including the mean for all students which was 12.1 out of 25 possible points.

Rasch calibration procedures calculate measurement fit statistics using a weighted infit and an unweighted outfit statistic (Table 3). The infit statistic is sensitive to midrange, organized misfit responses. The infit mean square statistic and standardized z scores (which provides a t -test of significance for the mean square values) indicate how accurate the measure is for a given student by assessing how far off the student is from the expected pattern of response. This is based on the odds for that student to respond in the same manner as others taking the assessment. An infit standardized z score greater than 2 indicates a less probable score reflecting too much noise (called underfit – student was unpredictable). Noise in this context refers to misinformation relative to measurement of student understanding. An infit z score less than -2 indicates the odds were too perfectly met (called overfit – student was too predictable), but is not considered a great threat to measurement. The outfit statistic is similar to the infit, but can better account for persons with unpredictable responses on the low end or high end of the scale. Outfit is more sensitive to random misfits such as test anxiety or loss of focus due to length of an assessment (cognitive fatigue).

The student infit mean was $Zstd = 0$, and the maximum value, and minimum value of all statistics were less than $Zstd = 2$, indicating no strong concerns about these parameters. The outfit statistic for maximum indicates potential for concern, though no person had corresponding infit indices above $Zstd = 2$, so all person data was retained.

Item measures. The items statistics provided by Rasch measurement can be interpreted in a similar way as the student statistics, except the focus moves from student performance

Table 3. Understanding items: student infit–outfit.

Student infit–outfit	QM Bugs II understanding items (Q1–Q25)	
	Infit <i>Zstd</i>	Outfit <i>Zstd</i>
Mean	0.0	0.1
Standard deviation	1.1	1.1
Max	1.7	3.1
Min	–2.9	–2.2

Table 4. Item summary for the QM BUGS II understanding items.

Item summary	QM BUGS II understanding items (Q1–Q25)	
	Total	Measure
Mean	32.0	0
Standard deviation	13.5	1.12
Max	55	2.88
Min	4	–1.93
Reliability		0.92
Separation		3.43

Table 5. Understanding items: item infit–outfit.

Item infit–outfit	QM Bugs II understanding items (Q1–Q25)	
	Infit <i>Zstd</i>	Outfit <i>Zstd</i>
Mean	0.0	0.1
Standard deviation	1.0	1.1
Max	2.0	2.3
Min	–2.7	–2.5

to item difficulty (Table 4). The total measure mean for items was calculated using measures by students on each individual item. While Rasch fit analysis indicated that there were concerns with some items, these items were not removed. Each item assesses a given characteristic of quantitative modelling understanding, so the research team effort is on identifying and revising problematic items, rather than removing them. All items had relatively good fit to the Rasch model, with only Q19 having borderline underfit (Outfit *Zstd* = 2.3).

Item infit and outfit means are summarized in Table 5. Infit and outfit mean, standard deviation, and maximum values were less than or near *Zstd* = 2, so they were not flagged as an overall concern. The QM BUGS II assessment yielded item reliability levels of 0.92 indicating the assessment showed strong evidence for reproducibility. Separation indices of 3.43 reflects at least three distinct levels of item difficulty.

Joint item-student analysis: Rasch ruler. The measures for student and item are jointly considered in Rasch measurement. One of the primary ways of viewing the interaction between student and items is the Rasch ruler, or Variable Map (Wilson, 2009; Wright & Stone, 1979), which places the students and item measures on the same scale graphically.

The Rasch ruler for the QM BUGS II assessment is provided in Figure 1. Person measures are plotted on the left; item measures on the right, where the mean (M), standard deviation (S), and two standard deviations (T) are shown. In Rasch measurement, item difficulty is based on the probability that a student will answer an item correctly. A person has a 50% chance of getting the items ‘correct’ (indicating the most ideal response) when those item measure values are the same as their person measure value. For example, on the assessment, those students who are at the mean score measure of 0 have a 50% chance of responding to item Q2 with the best answer. The items higher on the difficulty scale than the person measure are less likely to elicit correct responses by that individual. The higher the items are on the scale the more difficult they are for the student to answer correctly. Similarly, the lower the item is on the scale the easier it is for the student to respond correctly. Misfit, or unexpected responses, are flagged by Rasch fit statistics when a student correctly answers questions that are especially difficult for them (above their person measure) or miss items that are predicted to be especially easy for them (below their person measure).

The Rasch ruler (Figure 1) allows us to compare the distribution of students to the distribution of specified items. The student distribution was relatively well targeted with respect to the item distribution, indicating that overall the assessment was not too difficult for the students. For the QM BUGS II assessment understanding items, no person measures exceeded all items, though Q14 and Q11 were above all person measures.

Figure 2 provides a histogram of the Rasch ruler, which allows another visual representation of the overlap of students with items in aggregate. We are looking for positive overlap of student and item measures, which is largely evident in the histogram. Overlap of item and person measures helps maximize measurement accuracy and identifies whether the assessment is well suited, or targeted, to the ability level of the participants. We also see that the person measures are lower than two of the item measures. This lack of overlap tends to increase error for those item and person measures that do not overlap. A goal for revising the assessments will be to better align the overlap between students and items.

Rasch analysis of QM confidence items

An analysis similar to that conducted for the QM BUGS II understanding items was run for the QM confidence items separately. The assessment Likert scale items on QM confidence (Q26–Q36) allowed students to choose from a 4 level scale with 2 representing strongly disagree to 5 representing strongly agree. A rating of 1 was used for Not Applicable so this was not part of the ordinal scale, but a means for students to opt out for the item. Thus the minimum possible raw score on the assessments was 11 and the maximum raw score was $11 \times 4 = 44$.

A Rasch Rating Scale (Andrich, 1978) calibration was run on QM BUGS confidence rating items (Q26–Q36) to identify potential problematic items and students. The research team did not eliminate items from calibrations, so all 11 items were included in all analyses.

Reliability and validity indices for confidence items. Rasch calibration calculations of a person reliability index were conducted on the confidence items (Table 6). The QM BUGS II confidence items yielded person reliability levels of 0.88 indicating the assessment showed an acceptable level of reliability for person data. Separation indices of 2.65 reflects at least two distinctions in levels of students. The Total Score column provides a

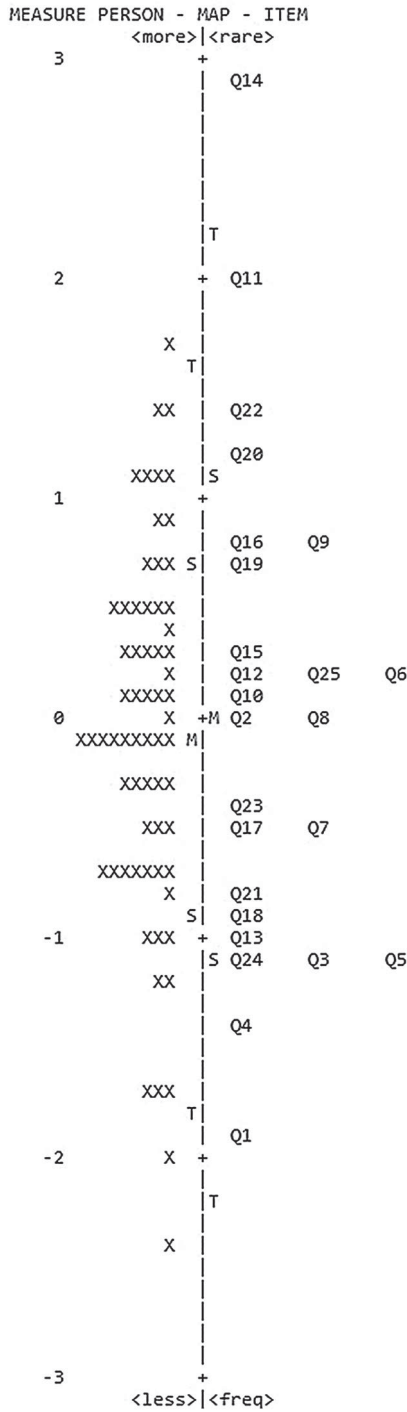


Figure 1. Variable map representation of the Rasch ruler for understanding items. Items are specified (right distribution) to support accurate evaluation of targeting to persons (left distribution).

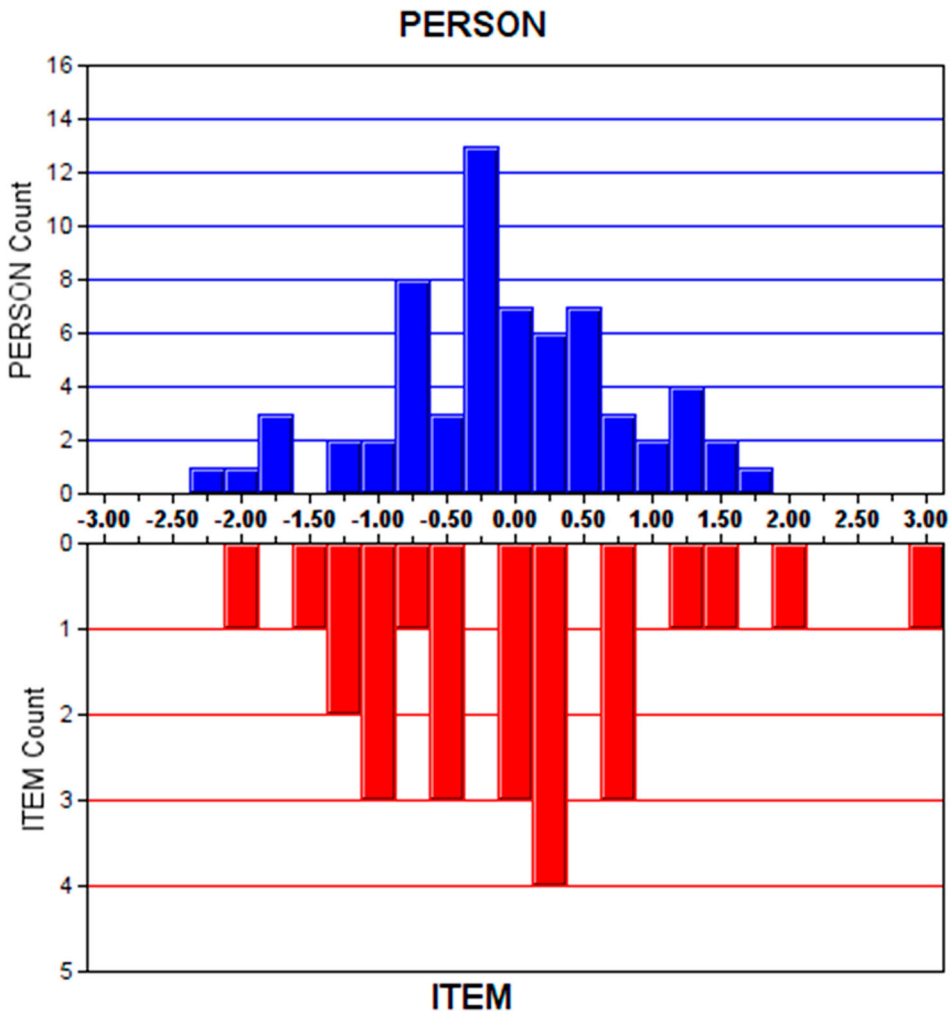


Figure 2. Histogram representation of the Rasch ruler for understanding items. Person measures of understanding can be compared with item measures of difficulty, given both measures are calibrated to the same Rasch scale.

student raw score summary, including that the mean for all students was 39.4 out of 44 possible points.

Person measures. Comparing raw score data, or percentages of those scored, can be problematic when using Likert choice scales because it incorrectly assumes that each ordinal level on the scale is an equivalent point toward the total, leaving a misleading suggestion that an average rating is equivalent to a measure. As with the multiple choice items, Rasch methods calculate a scale score or measure from the raw ratings which address this problem. Again, the measure is a standardized scaled score that is plotted on the Rasch ruler (Table 6). The mean measure of .22 indicates that students scored slightly above the item difficulty mean on the assessment.

Table 6. Person summary for the QM BUGS II confidence items.

Person summary	QM BUGS II confidence items (Q26–Q36)	
	Total raw score	Measure
Mean	39.1	.22
Standard Deviation	8.2	2.60
Max	53	6.45
Min	3	−7.42
Reliability	1.00	0.88
Separation		2.65

Table 7. Confidence items: student infit–outfit.

Student infit–outfit	QM Bugs II confidence items (Q26–Q36)	
	Infit	Outfit
Mean	−.3	−.3
Standard deviation	1.4	1.4
Max	2.8	2.8
Min	−4.3	−4.3

Table 8. Item summary for the QM BUGS II confidence items.

Item summary	QM BUGS II confidence items (Q26–Q36)	
	Total raw score	Measure
Mean	234.8	0
Standard deviation	8.4	0.47
Max	255	0.57
Min	222	−1.05
Reliability	N/A	0.60
Separation	N/A	1.23

Rasch calibration procedures calculated for confidence items provided measurement fit statistics using a weighted infit and an unweighted outfit statistic (Table 7). Person infit and outfit mean values were near the ideal of zero. The infit score maximum value was greater than 2, which indicates underfit, implying at least one student score was unpredictable for the maximum value. The outfit statistics is also a concern for the maximum score. These results call for review of individual fit scores for persons. Review of individual person fit statistics did not raise significant concern to justify removing student data, however.

Item measures. The QM BUGS II assessment yielded an item reliability level of 0.60 indicating the assessment showed moderate reliability (Table 8). Separation indices of 1.23 reflects little distinction in levels of items.

Item infit and outfit means for the confidence items are summarized in Table 9. Infit and outfit means were at or near the ideal of zero. All outfit values were also less than $Z_{std} = 2$, so evidence indicated that they corresponded with the Rasch model. Item fit analysis did not indicate that there were concerns with any items, so no items were removed. Category use (Figure 3) among the four levels was above the minimum of 10 necessary for meaningful analyses with the greatest use for level 4 (424 times) and level three (175 times). Level 2

Table 9. Confidence items: item infit–outfit.

Item infit–outfit	QM Bugs II confidence items (Q26–Q36)	
	Infit Zstd	Outfit Zstd
Mean	−.1	−.1
Standard deviation	.9	.6
Max	1.1	1.3
Min	−1.9	−1.1

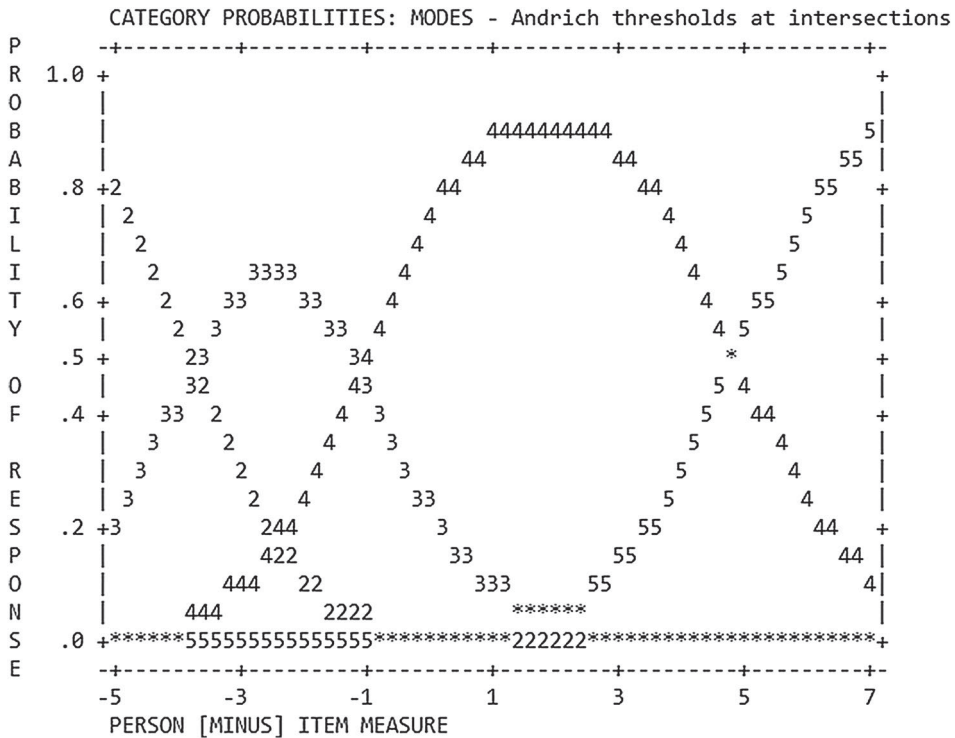


Figure 3. Rating scale category probability curves for confidence items indicate relative use and distinctions among the four levels.

(51 times) and level 5 (52 times) were nearly identical in category use. The opt-out choice of 1 (Not Applicable) was used 20 times, which suggests that its inclusion among the options may support accurate responding.

Polarity and item measure correlation. Rasch analysis provides a point measure correlation to examine polarity and correlation strength for item measures. All 11 items showed positive polarity with correlations between .71 and .80 indicating strong cohesion among confidence items.

Joint item-student analysis: Rasch ruler. The Rasch ruler for the QM BUGS II confidence assessment items is provided in Figure 4. The student distribution is higher with respect to the item distribution, indicating that overall the assessment item statements were not

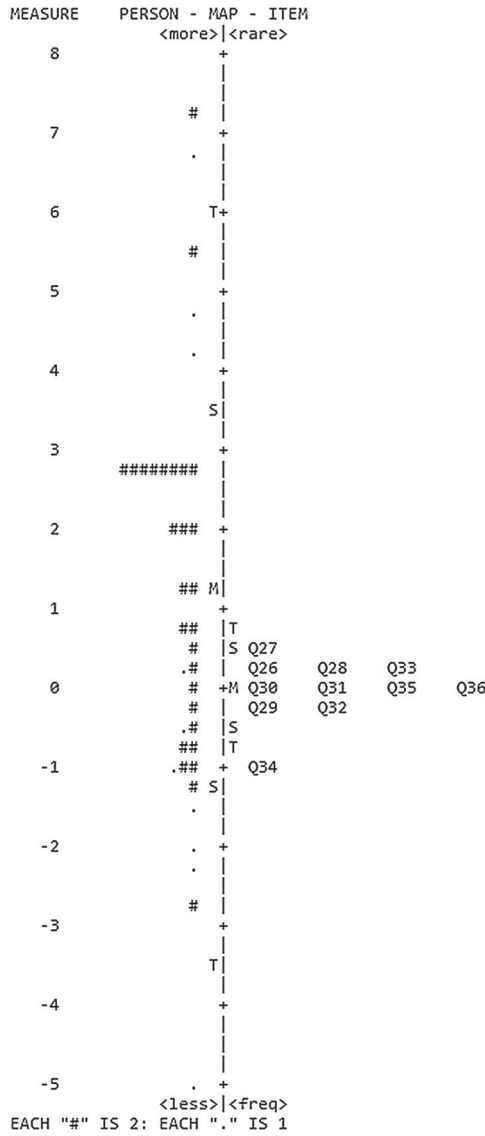


Figure 4. Variable map representation of the Rasch ruler for confidence items. Items are specified (right distribution) to support accurate evaluation of targeting to persons (left distribution).

very difficult for the students to select. For the QM BUGS II confidence assessment items many of the person measures exceeded all items. So the targeting of items to persons for the confidence items was not ideal for this sample. Note how all the items are clustered except for Q34 on confidence in making predictions from a model.

Figure 5 provides a histogram of the Rasch Ruler, which provide an alternate visual representation of the overlap of students with items. We are looking for positive overlap of student and item measures, which is evident only at the center of the histogram but not at the two tails. Overlap of item and person measures helps maximize measurement accuracy

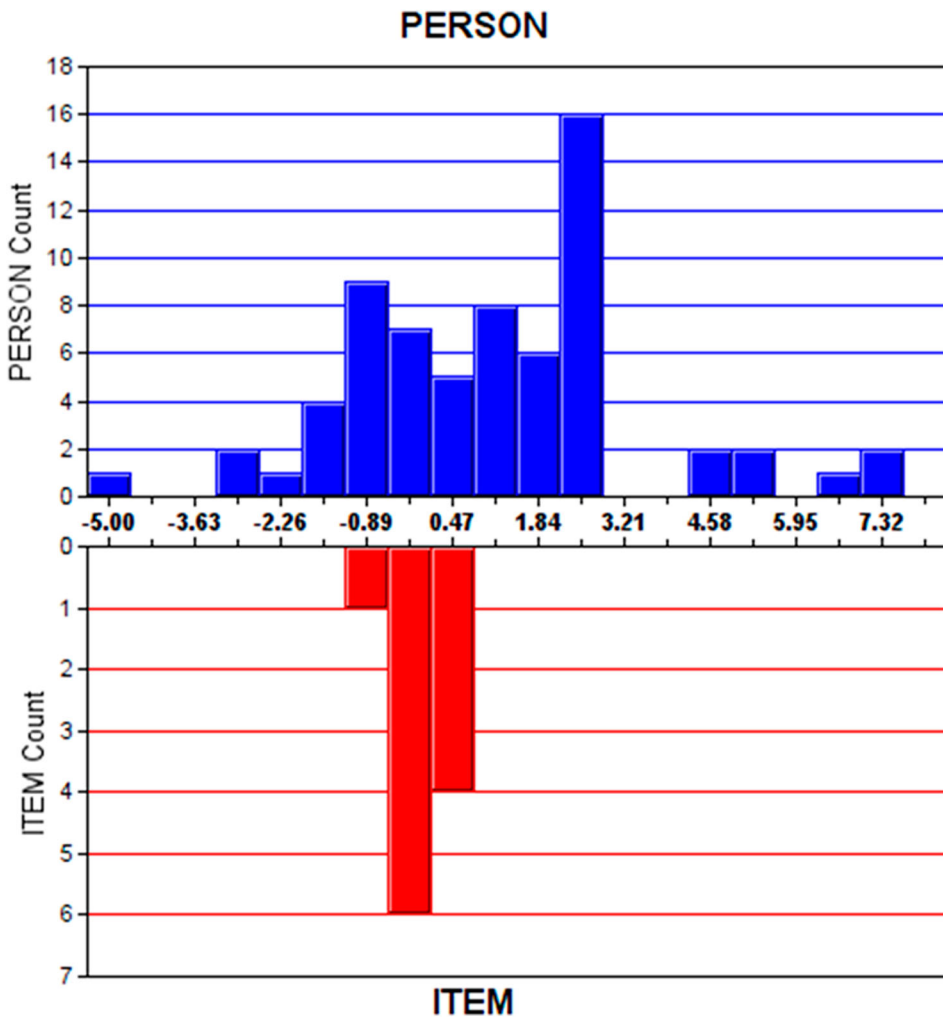


Figure 5. Histogram representation of the Rasch ruler for confidence items. Person measures of confidence can be compared with item measures of difficulty to endorse, given both measures are calibrated to the same Rasch scale.

and identifies whether the assessment is well suited, or targeted, to the ability level of the participants. This lack of overlap on each end tends to increase error for those item and person measures that do not overlap. A goal for revising the assessments will be to better align the overlap between student and items.

Discussion

To answer the question of what impact increasing the amount of modelling in an undergraduate biology course has on either disciplinary knowledge (Objective 1) or quantitative modelling of biology phenomena (Objective 2), we first focused on developing a valid diagnostic assessment of students' quantitative modelling abilities and confidence in biology (Objective 3). However, piloting the QM BUGS II assessment did provide some

insights into student QR ability in low to moderate intensiveness integration of QR into the courses. This could be interpreted as ‘What is the current state of students QR ability in undergraduate biology courses?’

Implications of findings

In this study, the QM BUGS II assessment’s understanding measure was moderately reliable and weak in distinction levels, relative to the grouping of students into weaker versus stronger understanding levels. In addition, the Rasch ruler indicates the assessment of understanding was not too difficult for the students, but there is room to improve alignment of the assessment with students. These limitations call for further analysis of the assessment items such as the planned qualitative analysis of student interactions with items to support a more thorough interpretation of QM BUGS III.

On the other hand, the assessment’s Confidence measure showed acceptable reliability and it separated students into two distinct groups of weaker confidence and stronger confidence. Yet the relatively high ratings on confidence may reflect a lack of alignment of the confidence items with specific topics understood by these students that we should consider as confidence items are also carefully reviewed.

Some specific items to revise were identified by the Rasch analysis. The descriptive statistics and Rasch analysis identify Q14 and Q22 as problematic. Descriptive statistics also indicate the performance on Q11 was extremely poor, while Rasch analysis identified Q19 as having low item measure correlation. The following are possible explanations for the low correlation. Q11 includes more extensive text than other items, which was considered necessary to provide enough theoretical background for students to determine a model mechanistically. Students may perform worse on this item due to reading comprehension or a memory load burden that may be required in thoughtfully responding to it. Q14 is a negatively phrased question requiring students to identify the answer which is NOT appropriate for empirically testing a model. The negative statement of the question may also be affecting cognitive load during interpretation, resulting in poor student performance. Q22 asks students to identify which answers are qualities (plural) of a model, but only one answer is correct. Students may err due to considering there is more than one correct response. Although these explanations have yet to be evaluated through revision and testing, they represent distinct possibilities whereby item demands may undermine the functioning of each respective item as a means to assess the respective type of understanding.

Revisions of the QM BUGS II assessment to address the issues outlined above were completed in fall 2018, resulting in QM BUGS III. Questions were vetted by biology researchers with expertise in quantitative reasoning in biology. This process addressed content and face validity for the assessment. This assessment was administered at the end of the fall 2018 semester. Rasch analysis will be conducted on the assessment data collected to determine reliability and validity evidence following this revision. If warranted by improved psychometric characteristics of the instrument a report on student quantitative modelling in biology ability and confidence will be an outcome of this analysis. Further revisions of the assessment will be made based on the Rasch analysis. A qualitative analysis of student interactions with items is planned to support future analysis of QM BUGS III. The instrument will be made available through the QUBES project portal.

Implications for educators

With the push in undergraduate biology education towards emphasis on core competencies, including modelling, simulations and quantitative reasoning, biology instructors must consider how and when their students develop these abilities (AAAS, 2011). Students who perform poorly on the QA and QI portions of the assessment may be at a disadvantage when confronted with the increasingly quantitative dimensions of biology curricula. Additionally, biology instructors need to determine whether students recognize the nature and purpose of modelling (Papaevripidou & Zacharia, 2015). This assessment will serve as a reference for faculty that seek to integrate quantitative modelling into biology courses as they consider the breadth of skills that underlie quantitative modelling in biology.

An implication of this investigation of quantitative modelling and its findings for biologist-educators is the potential for adaptability of the process steps described within this study for critically evaluating assessment items written or included within quizzes, exams, tests, or self-report ratings administered to biology students. While domain expertise of educators in the biological sciences, along with item-development experience, the revision process, and the application of common sense can each contribute to an educator's development of effective assessment items, our findings presented here illustrate some example outcomes indicating why it is best not to rest on one's intuitive sense of satisfaction with assessment items created or adapted. For instance, some items may contribute to effective measurement of understanding on a particular biological topic area, but other items developed for that same topic may not contribute well to the measurement, and thereby degrade the overall clarity of interpretations that result from data. Furthermore, characteristics such as student ability, motivation, and prior knowledge can all contribute to the degree of item-person-targeting, as demonstrated with the use of Rasch model calibration and variable map construction. Other factors including instructional effectiveness and curriculum content choices may likewise influence reliability and validity findings of assessment outcomes. Thus, the procedures described within this study can be incorporated into a departmental or an individual educator's toolkit as a means of continuing to improve upon the assessment process and the potential benefits derived from the assessment process for the students and the educator. In particular, we encourage the construction of Rasch scales, rather than relying solely on raw-score sums and percentages; the use of variable maps for item-person targeting analyses; the use of item-measure correlation statistics to examine polarity and cohesiveness; and the use of fit statistics for both person measures and item measures to identify whether items generally appear to be measuring a common construct or content area. Fortunately, both specialized software (e.g. Winsteps) and larger statistical packages and languages (e.g. R) are increasingly incorporating Rasch functionality, making the use of these analytic tools more feasible for educators and researchers.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Science Foundation under the Supporting Faculty in Quantitative Undergraduate Biology Education and Synthesis (QUBES) [grant number 1446258].

Data availability statement

Access to data set associated with paper is not provided.

ORCID

Joseph Dauer  <http://orcid.org/0000-0002-8971-0441>

References

- Aikens, M. L., & Dolan, E. L. (2014). Teaching quantitative biology: Goals, assessments, and resources. *Molecular Biology of the Cell*, 25(22), 3478–3481.
- Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education*, 33, 131–152. doi:10.1016/S0360-1315(99)00029-9
- American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education: A call to action*. Retrieved from <http://visionandchange.org/files/2011/03/Revised-Vision-and-Change-Final-Report.pdf>
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Association of American Medical Colleges, and Howard Hughes Medical Institute. (2014). *Scientific foundations for future physicians: Report of the AAMC-HHMI committee*. Washington, DC.
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Routledge.
- Dahlquist, K. D., Aikens, M. L., Dauer, J. T., Donovan, S. S., Eaton, C. D., Highlander, H. C., . . . Schugart, R. C. (2018). *Peeking into the black box: Models as an effective epistemic tool for building student disciplinary knowledge and practices/skills*. Report on QUBES project.
- De Ayala, R. J. (2011). *The theory and practice of item response theory*. New York, NY: The Guilford Press.
- Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academies Press.
- Eaton, C. D., Highlander, H. C., Dahlquist, K. D., LaMar, M. D., Ledder, G., & Schugart, R. C. (2017). A ‘rule of five’ framework for models and modeling to unify mathematicians and biologists and improve student learning. Revised version submitted to the journal *PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies* on April 28, 2017, available from the arXiv preprint server. Retrieved from <https://arxiv.org/abs/1607.02165v2>
- Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.
- Garfunkel, S., & Montgomery, M. (2016). *Guidelines for assessment and instruction in mathematical modeling education (GAIMME)*. Boston, MA: Consortium for Mathematics and Its Applications (COMAP)/Society for Industrial and Applied Mathematics (SIAM).
- Krajcik, J., & Merritt, J. (2012). Engaging students in scientific practices: What does constructing and revising models look like in the science classroom? *The Science Teacher*, 79(3), 38.
- Lehrer, R., & Schauble, L. (2000). Developing model-based reasoning in mathematics and science. *Journal of Applied Developmental Psychology*, 21(1), 39–48.
- Lehrer, R., & Schauble, L. (2005). Developing modeling and argument in the elementary grades. In T. A. Romberg, T. P. Carpenter, & F. Dremock (Eds.), *Understanding mathematics and science matters* (pp. 29–53). New York, NY: Routledge.
- Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8), 816–834.
- Linacre, J. M. (2012). *Winsteps*[®] (Version 3.74.0) [Computer Software]. Retrieved from <http://www.winsteps.com/>
- Linacre, J. M. (2014). *Winsteps*[®] Rasch measurement computer program user’s guide. Beaverton, OR: Winsteps.com.

- Louca, L. T., & Zacharias, Z. C. (2012). Modeling-based learning in science education: Cognitive, metacognitive, social, material and epistemological contributions. *Educational Review*, 64(4), 471–492.
- Magnani, L., Nersessian, N. J., & Pizzi, C. (Eds.). (2012). *Logical and computational aspects of model-based reasoning* (Vol. 25). Dordrecht: Kluwer Academic Publishers.
- Mayes, R., Forrester, J., Christus, J. S., Peterson, F., & Walker, R. (2014). Quantitative reasoning learning progression: The matrix. *Numeracy*, 7(2), 1–20.
- Mayes, R., Forrester, J., Christus, J., Yestness, N., Peterson, F., & Bonilla, R. (2013). Quantitative reasoning in environmental science: A learning progression. *International Journal of Science Education*, 36(4), 635–658.
- Metcalfe, S. J., Krajcik, J., & Soloway, E. (2000). Model-it: A design retrospective. In M. Jacobson & R. B. Kozma (Eds.), *Innovations in science and mathematics education: Advanced designs for technologies in learning* (pp. 77–116). Mahwah, NJ: Lawrence Erlbaum Associates.
- National Institute of Biomedical Imaging and Bioengineering. (2018). Retrieved from <https://www.nibib.nih.gov/science-education/science-topics/computational-modeling>
- NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press NRC.
- Nicolaou, C. T., & Constantinou, C. P. (2014). Assessment of the modelling competence: A systematic review and synthesis of empirical research. *Educational Research Review*, 13, 52–73.
- Oh, P. S., & Oh, S. J. (2010). What teachers of science need to know about models: An overview. *International Journal of Science Education*, 32, 1–29.
- Paivio, A. (1990). *Mental representations: A dual coding approach*. New York, NY: Oxford University Press.
- Papaevripidou, M., Constantinou, C. P., & Zacharia, Z. C. (2007). Modeling complex marine ecosystems: An investigation of two teaching approaches with fifth graders. *Journal of Computer Assisted Learning*, 23(2), 145–157. doi:10.1111/j.1365-2729.2006.00217.x
- Papaevripidou, M., & Zacharia, Z. C. (2015). Examining how students' knowledge of the subject domain affects their processing of modeling in a computer programming environment. *Journal of Computers in Education*, 2(3), 251–282.
- Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. *Journal of Research in Science Teaching*, 48, 486–511.
- QUBES. (2019). Retrieved from <https://qubeshub.org/>
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. Expanded edition (1980). Chicago: University of Chicago Press.
- Robeva, R. (Ed.). (2015). *Algebraic and discrete mathematical methods for modern biology*. Boston, MA: Academic Press.
- Schwarz, C. V., Gunckel, K. L., Smith, E. L., Covitt, B. A., Bae, M., Enfield, M., & Tsurusaki, B. K. (2008). Helping elementary preservice teachers learn to use curriculum materials for effective science teaching. *Science Education*, 92(2), 345–377.
- Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Archer, A., Fortus, D., & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learnings. *Journal of Research in Science Teaching*, 46, 632–654.
- Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and Instruction*, 23(2), 165–205. doi:10.1207/s1532690xci2302_1
- Sins, P., Savelsbergh, E. R., & van Joolingen, W. R. (2005). The difficult process of scientific modelling: An analysis of novices' reasoning during computer-based modelling. *International Journal of Science Education*, 27(14), 1695–1721.
- Snir, J., Smith, C. L., & Raz, G. (2003). Linking phenomena with competing underlying models: A software tool for introducing students to the particulate model of matter. *Science Education*, 87(6), 794–830.
- Stieff, M. (2017). Drawing for promoting learning and engagement with dynamic visualizations. In *Learning from dynamic visualization* (pp. 333–356). Cham: Springer.

- Svoboda, J., & Passmore, C. (2011). The strategies of modeling in biology education. *Science & Education*, 22, 119–142. doi:10.1007/s11191-011-9425-5
- Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46, 716–730.
- Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education*, 92, 941–967. doi:10.1002/sc.20259
- Winsberg, E. (2003). Simulated experiments: Methodology for a virtual world. *Philosophy of Science*, 70, 105–125.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Appendices

Appendix A. Modelling framework (superscripts provide key for cross modelling elements relationships)

Modelling framework	Modelling element	Sub-element	Q#	Description
Model formulation	Analyse ¹		Q1	Decompose phenomena into quantifiable variables; explain an anchoring phenomena - introduce driving questions and phenomena for a particular concept (Schwarz)
	Inductive Reasoning ² Quantitative Act	Variable Quantification ¹	Q2	Hypothesize how elements interact conceptually and quantitatively
			Q3 Q4	Mental construct for object within context including both attributes and measure (Thompson); capacity to communicate quantitative account of solution, decision, course of action within context
		Variation ²	Q2	Reason about covariation of variables; comparing, contrasting, relating variables in the context of problem
		Quantitative Literacy ²	Q5	Reasons with quantities to explain relationships between variables; proportional reasoning, numerical reasoning; extend to algebraic and higher math reasoning
		Context	Q6	Situative view of QR within a community of practice (Shavelson); solves ill-defined problems in socio-political contexts using ad-hoc methods; informal reasoning within science context (Steen & Madison; Sadler & Zeidler)
	Quantitative Modelling	Create Model ³	Q7	Ability to create a model representing a context and apply it within context; use variety of quantitative methods to construct model including least squares, linearization, normal distribution, logarithmic, logistic growth, multivariate, simulation models
			Q9	
			Q10	
			Q12	Conduct statistical inference to test hypothesis (Duschl)
		Refine Model ⁴	Q13	Extend model to new situation; test and refine a model for internal consistency and coherence to evaluate scientific evidence, explanations, and results; (Duschl)
		Model Reasoning ⁵	Q20–Q24	Construct and use models spontaneously to assist own thinking, predict behaviour in real-world, generate new questions about phenomena (Schwarz)
	Phenomenological Model ³		Q8	Models only represent the observable properties of the phenomenon, refrain from including the actual underlying mechanism (Louca and Zacharia). Model foregoes any attempt to explain why the variables interact the way they do, and simply attempts to describe the relationship.
Mechanistic Model ³		Q11	Model assumes complex system can be understood by examining workings of its individual parts and manner in which they are coupled. Mechanistic models typically have a tangible, physical aspect, in that system components are real, solid and visible.	
Model Deployment	Test Model ⁴		Q13	Investigate the phenomena and the interactions with model
	Model Evaluation ⁴		Q13	Assess degree of fit and ways to change model
	Model Application ⁶		Q18	Use the model to predict or explain other phenomena
	Model Validation	Empirical Assessment	Q14	Model can explain all of the data and predict future experiments. Assess whether a model can explain all of the data at hand and predict the results of future experiments.

(continued)



(Appendix A. Continued)

Modelling framework	Modelling element	Sub-element	Q#	Description
	Model Comparison ⁷	Conceptual Assessment	Q15	Evaluate how well a model fits with other accepted models and knowledge
	Quantitative Interpretation	Trends ⁶	Q16	Determine multiple types of trends including linear, power, and exponential trends; recognize and provide quantitative explanations of trends in model representation within context of problem
		Predictions ⁶	Q18	Makes predictions using covariation and provide quantitative account applied within context of problem
		Translation ⁷	Q17	Translates between models; challenges quantitative variation between models as estimates or due to measurement error; identifies best model representing a context
		Revision ⁷	Q19	Revise models theoretically without data, evaluate competing models for possible combination (Schwarz). Models are continually revised to probe new phenomena and account for new data.
Modelling Reasoning	Metacognitive Knowledge		Q25	Self-regulated learner explicitly identifies and describes the major steps of modelling process (Papaevripidou and Zacharia)
	Meta-modelling Knowledge	Nature of Models ⁸	Q20–Q23	Epistemic knowledge corresponding to understanding of nature of models (Schwarz and White; Oh and Oh)
		Purpose/Utility of Model ⁹	Q24	Appreciation of the purpose and utility of models (Schwarz and White; Oh and Oh)
	Model-based Reasoning ⁵	Model as Ideas ⁸	Q23	Models are ideas not physical objects.
		Multiple Representations ⁸	Q20	Models are communicated through drawings, graphs, equations, three-dimensional structures or words. The representations are distinct from the underlying model they purport to explain.
			Acceptability ⁸	Q22 Q23
	Uniqueness ⁸		Q23	Not always possible or even desirable to exclude all but one model. Different models may account for different aspects of a phenomena.
		Development ⁸	Q23	An experiment and observations inform the development of a model.
	Empirical or Theoretical Objects ⁸		Q21	Models are constituted by a set of objects which may be empirical (genes and alleles in meiotic model) or theoretical objects.
		Processes ⁸	Q21 Q23	Models are constituted of processes in which objects participate.
	Scientific Model ⁹		Q20	An idea or set of ideas that explains what causes a particular phenomenon in nature (Modelling for Understanding in Science Education – MUSE)
	Application ⁹		Q24	Model applied to explain reality, make predictions , assess for how well it explains real-world phenomena.
	Guide Future Work ⁹		Q24	Models influence and constrain questions scientists ask about natural world and types of evidence they seek.

Appendix B. Sample items from QM BUGS II

QM BUGS II

QA Quantitative Literacy

5. Stomatal conductance R_{vs} is the rate of passage of water vapor exiting through the stomata (small pores) of a leaf. A steady state porometer (Figure 3) is an instrument that measures stomatal conductance by clamping it to the leaf surface, then computing the vapor flux between two locations on the diffusion path. The ratio of the change between vapor concentration at the leaf C_{vL} and the concentration at the first sensor C_{v1} with the combined stomatal resistance R_{vs} and resistance at the first sensor R_1 is used in the vapor flux computation. Which of the following expressions represents this ratio? (QA Quantitative Literacy - a)

- $\frac{C_{vL} - C_{v1}}{R_{vs} + R_1}$
- $(C_{vL} - C_{v1})(R_{vs} + R_1)$
- $\frac{C_{vL} - C_{v2}}{R_{vs} + R_2}$
- $\frac{C_{vL} - C_{v1}}{R_{vs} + R_1}$
- $\frac{C_{vL} + C_{v1}}{R_{vs} - R_1}$



Figure 3

QM BUGS II

QM Phenomenological Model

8. The biologist is working on building a model of Transpiration Rate (y-axis) by Temperature (x-axis) from the data in Table 3. Biologists created a scatterplot of the data (Figure 5) and fit different models to these data. Select the best model for the data. (QM Create Model - phenomenological - b)

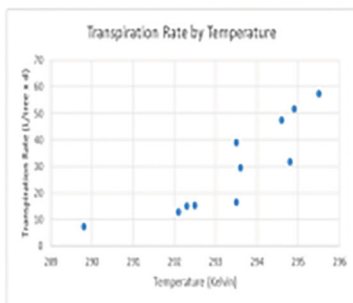
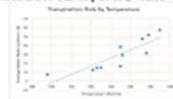


Figure 5

Table 3	
T (K)	TR (L/tree x d)
289.8	7.3
292.1	12.8
292.3	15.1
292.5	15.3
293.5	16.6
293.5	38.9
293.6	29.4
294.6	47.5
294.8	31.7
294.9	51.8
295.5	57.6
296.4	10.9

- a. A linear model $y = ax + b$ that indicates a constant rate of increase in transpiration rate with an increase in temperature, with $R^2 = 0.747$.



- b. A quadratic model $y = ax^2 + bx + c$ that indicates a nonlinear increasing transpiration rate with increase in temperature, with $R^2 = 0.834$.



- c. A linear model $y = ax + b$ that indicates a constant rate of increase in transpiration rate with an increase in temperature, has $R^2 = 0.508$, and which passes through the 0-K point and contains as many points of the data set as possible.



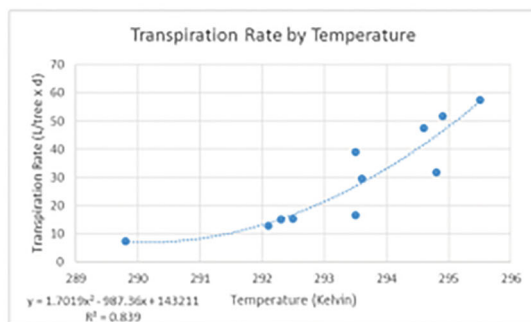
- d. A quadratic model $y = ax^2 + bx + c$ that indicates a nonlinear increasing transpiration rate where the rate of increase is decreasing with increase in temperature, $R^2 = 0.834$.



QM BUGS II

QI Model Application - Prediction

18. Use the graph or equation model in Figure 11 to predict what happens to transpiration rate if the temperature is 296 K? Select the best answer. (Model Application QI Prediction - c)



- Data was only collected for 289.8 to 295.5 degrees Kelvin, so you can't make a prediction for transpiration rate above 295.5 degree Kelvin.
- Temperature continues to increase across the graph but cannot make an estimate beyond the graph.
- Transpiration rate is increasing as temperature increases, so at 296 degree Kelvin the transpiration rate is approximately 65 L/(tree x d).
- After 295.5 degrees Kelvin the transpiration rate remains at a constant value of approximately 56.5 L/(tree x d).
- Transpiration rate decreases after 295.5 degrees Kelvin and so is less than 56.5 L/(tree x d).

QM BUGS II

MM Purpose and Utility of Models

24. Which of the following best represents the purpose and utility of a model? (Meta-modeling, Purpose and Utility of Models - a)
- Models influence and constrain the kinds of research questions that guide future work of scientists and are not limited to the study from which they arise.
 - Models are theoretical and are not meant to explain reality.
 - Models are based on collected data and are not meant to be used to make predictions about future events.
 - Models stand on their own and do not have to be assessed for how well they explain real-world phenomena.
 - Models do not influence or constrain the kinds of questions that guide future work of scientists.

QM Confidence Item Sample

Relative to developing models of plant transpiration which you explored in the first part of this assessment, rate your level of confidence for each question below. (QA Variable and Hypothesis)

26. Working from verbal descriptions of plant transpiration, I was able to determine the variables needed for the model, identifying both properties of the variables useful in building a model and an appropriate unit of measure for the variables. (QA Variable quantification)

Not Applicable	Strongly Disagree	Disagree	Agree	Strongly Agree
NA	1	2	3	4

30. Conducting a formal statistical test to determine significance of a hypothesis. (QM Statistics)

Not Applicable	Strongly Disagree	Disagree	Agree	Strongly Agree
NA	1	2	3	4

35. I was comfortable determining trends in the transpiration data and defending those trends using biological and mathematical arguments. (QI trends)

Not Applicable	Strongly Disagree	Disagree	Agree	Strongly Agree
NA	1	2	3	4