2021

Assessing Quantitative Modelling Practices, Metamodelling, and Capability Confidence of Biology Undergraduate Students

Joseph Dauer
University of Nebraska-Lincoln, joseph.dauer@unl.edu

Robert Mayes
Georgia Southern University, rmayes@georgiasouthern.edu

Kent Rittschof
Georgia Southern University, kent_r@georgiasouthern.edu

Bryon Gallant
Georgia Southern University

Dauer, Joseph; Mayes, Robert; Rittschof, Kent; and Gallant, Bryon, "Assessing Quantitative Modelling Practices, Metamodelling, and Capability Confidence of Biology Undergraduate Students" (2021). *Papers in Natural Resources*. 1469.
https://digitalcommons.unl.edu/natrespapers/1469

Assessing quantitative modelling practices, metamodelling, and capability confidence of biology undergraduate students

Joseph Dauer,[1] Robert Mayes,[2]
Kent Rittschof,[3] and Bryon Gallant[4]

1 School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE, USA
2 Middle Grades and Secondary Education, Georgia Southern University, Statesboro, GA, USA
3 Curriculum, Foundations, and Reading, Georgia Southern University, Statesboro, GA, USA
4 Clinical Psychology, Georgia Southern University, Statesboro, GA, USA

Correspondence — Joseph Dauer joseph.dauer@unl.edu School of Natural Resources, University of Nebraska-Lincoln, 3310 Holdrege St., 524 Hardin Hall, Lincoln, NE 68583, USA

ORCID
Joseph Dauer http://orcid.org/0000-0002-8971-0441
Robert Mayes http://orcid.org/0000-0003-2726-2409
Kent Rittschof http://orcid.org/0000-0003-3213-7266

Abstract
Quantitative modelling plays an important role as biology increasingly deals with big data sets, relies on modelling to understand system dynamics, makes predictions about impacts of changes, and revises our understanding of system interactions. An assessment of quantitative modelling in biology was administered to students (n = 612) in undergraduate biology courses at two universities to provide a picture

Published in *International Journal of Science Education* 2021, vol. 43, no. 10, pp. 1685–1707.
doi:10.1080/09500693.2021.1928325

Submitted 19 June 2020; accepted 5 May 2021; published 23 May 2021.

of student ability in quantitative reasoning within biology and to determine how capable those students felt about this ability. A Rasch analysis was used to construct linear measures and provide validity evidence for the assessment and to examine item statistics on the same scale as student ability measures. Students overall had greater ability in quantitative literacy than in quantitative interpretation of models or modelling. There was no effect of class standing (Freshmen, Sophomore, etc.) on student performance. The assessment showed that students who participated felt confidence in their ability to quantitatively model biological phenomena, even while their performance on ability questions were low. Collectively modelling practices were correlated with students' metamodelling knowledge and not correlated with students' modelling capability confidence. Biology instructors who incorporate the process of modelling into their courses may see improved abilities of students to perform on quantitative modelling tasks.

Keywords: Models and modelling, Rasch process, higher education

Introduction

National reports have repeatedly called for an increased emphasis on modelling in science, technology, engineering, and mathematics (STEM) education (AAAS, 2011; Association of American Medical Colleges & Howard Hughes Medical Institute, 2009; Garfunkel & Montgomery, 2016; National Research Council (NRC), 2003; NGSS Lead States, 2013). Modelling takes on many forms, including experiential (physical manipulatives), visual, verbal (qualitative discourse), numerical (quantitative data), or symbolic quantitative models (Diaz Eaton et al., 2019). Our working definition is that a model is a simplified representation of real-world objects and their mechanistic or functional relationships, constructed for a purpose, such as understanding or making predictions about a real-world phenomenon (Diaz Eaton et al., 2019).

Students who are given the opportunity to develop, refine, and test 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).

Ongoing efforts to cultivate authentic science practices in students have focused on developing their model-based reasoning skills and metamodelling abilities by engaging them in the modelling process (Hester et al., 2018; Schuchardt & Schunn, 2016) with an end goal of being able to generate 'defensible explanations for the way the natural world works' (Windschitl et al., 2008, p. 15). Developing these defensible explanations through modelling builds deeper understanding of biology (Louca & Zacharia, 2019). In the field of biology, quantitative models have taken on a major role given the explosion of both experimental data related to complex global problems and the software and inexpensive hardware that permit data analysis and simulation (Diaz Eaton et al., 2020; Li et al., 2010). Quantitative Modelling (QM), which we define as mathematically relating model components to describe system dynamics, is a critical skill for biology students and requires instructors thoughtfully integrate key quantitative dimensions into biology teaching (Mayes et al., 2014).

While there have been numerous calls to develop curriculum and provide professional development at the interface of mathematics and biology (Diaz Eaton et al., 2020), important questions remain about the degree to which QM is required to impact student understanding of biology. This paper provides insight into the current state of students' QM ability and confidence. A quantitative modelling in biology assessment (QM BUGS III) was administered to students in undergraduate biology courses and provides a picture of student ability in quantitative modelling within biology and how capable students feel about this ability.

Modelling practices

Despite its importance in promoting new knowledge, there is a deficit of research on the cognitive components, metacognitive processes, and impacts of modelling across STEM, especially for undergraduate students (Louca & Zacharia, 2019; Seel, 2017). The QM BUGS assessment used in this study attempts to account for cognitive and metacognitive processes used in modelling. While this research does not speak directly to impact of modelling in undergraduate biology, it does address the current state of students' QM ability and confidence. Modelling-based learning (MbL), on which the QM BUGS project is based, is a theoretical framework whereby learning takes place via student construction of models as representations of physical phenomena at the same time they are aware of the nature and purpose of those models (Louca & Zacharia, 2019; Schwarz & White, 2005; Windschitl et al., 2008). Model-based learning (MbL) uses student creation of models of real-world phenomena to develop deeper conceptual understanding (Gobert & Buckley, 2000). In science education, the MbL approach is grounded in inquiry, constructivism, and constructionism traditions (Schwarz et al., 2009) and is used to promote scientific literacy and authentic scientific inquiry (Acher et al., 2007). The construction and refinement of models has been shown to improve conceptual understanding, operational understanding of the nature of science, procedural and reasoning skills (Harrison & Treagust, 2000; Tsui & Treagust, 2013), science communication (Penner, 2000), peer collaboration (King et al., 2019), and metacognition (Jonassen et al., 2005). Moreover, modelling is an interdisciplinary skill, where modelling improvement in one discipline can transfer to other disciplines (Bamberger & Davis, 2013).

There is a pressing need to research the relationship between the epistemological aspect of MbL in science and the metacognitive processes students engage in through MbL. The epistemological knowledge that Schwarz and White (2005) identified as central to MbL can improve understanding of practices like predicting, observing, and explaining phenomena (Sins et al., 2005) and the ability to make mechanistic explanations (explaining phenomena in purely physical or deterministic terms) (Fretz et al., 2002; Louca et al., 2011). Importantly, modelling metacognition enhances students' abilities to regulate their own learning with models (Papaevripidou et al., 2007). As students gain awareness of where they are relative to a learning progression of modelling (Schwarz et al., 2009) they can be more aware of how they are using models to address biological problems. QM BUGS assesses the current state of students' awareness of using models.

Modelling can take on many forms, including both qualitative and quantitative. Our focus is on Quantitative Reasoning (QR) based on a framework developed by Mayes et al. (2014) that proposes three elements of QR. (1) Quantitative Act (QA) which is quantifying a problem by conceptualizing the focal object and assigning units of measure to its attributes (Thompson, 2011). (2) Quantitative Modelling (QM) which is developing and revising models to explain phenomena related to the object and its attributes (Schwarz et al., 2009). (3) Quantitative Interpretation (QI) which is using models to make predictions (Gilbert, 1991; Koponen, 2007; Sensevy et al., 2008). The use of models in this framework requires students to reason with and about models, which is termed meta-modelling (Papaevripidou et al., 2007; Svoboda & Passmore, 2013). Meta-modelling includes understanding the nature of models and a models utility and purpose (Papaevripidou & Zacharia, 2015). In addition, our observations of undergraduate biology courses for this study, as well as conversations with faculty teaching the courses, indicated that students were not confident about implementing quantitative approaches within a biology context. This apparent lack of confidence interferes with students engaging in quantitative reasoning. We wanted to know more about the level of confidence in QR capabilities that the students possessed. The above led to the inclusion of three main constructs in our study: Modelling Practices, Meta-modelling, and Quantitative biology capability confidence. Research on student learning in quantitative biology courses has focused on improving students' numeracy skills, graphical data interpretation, and inferences from mathematical models (Hoffman et al., 2016; Speth et al., 2010). To date, two instruments have been developed to assess undergraduate biology students' quantitative literacy and interpretation of models (Deane et al., 2016; Stanhope et al., 2017), and a third instrument assesses biology majors' calculus comprehension (Taylor et al., 2020). There is a gap in the research concerning students' cognitive and metacognitive modelling abilities, student's confidence in applying QR in biology contexts, and the practices utilized by students while they build and revise quantitative models in undergraduate biology courses. We have developed an assessment of abilities and confidence to create and apply models in biology employing pre-calculus mathematics.

Assessment development

The first version of the Quantitative Modelling Biology Undergraduate Student Assessment (QM BUGS I) was developed by two experts in biology modelling and mathematical modelling (JD and RM) who have researched and taught at the intersection of their disciplines and had its foundation in elements common to many modelling frameworks (Duschl et al., 2007; Lehrer & Schauble, 2006; Louca & Zacharia, 2012; Mayes et al., 2014; Oh & Oh, 2011; Papaevripidou & Zacharia, 2015; Pluta et al., 2011; Schwarz et al., 2009; Schwarz & White, 2005). Following a Rasch analysis of QM BUGS I, a revised QM BUGS

II consisted of five subsections: 25 multiple choice questions addressing four subcategories within quantitative modelling understanding (modelling practices (MP) which includes quantitative act (QA), quantitative interpretation (QI), and quantitative modelling (QM) abilities); metamodelling knowledge (MMK); and 11 Likert questions on a 4-level scale addressing student quantitative biology capability confidence (QBCC). Confidence self-ratings addressed each of the QR abilities of QA, QM and QI. An extensive Rasch analysis was conducted on QM BUGS II to identify further revisions for the current QM BUGS III version (Mayes et al., 2019). QM BUGS III (assessment can be viewed at https://doi.org/10.32873/unl.dr.20201008) consists of 38 questions: 26 multiple choice questions addressing modelling and metamodelling, and 12 Likert questions addressing QBCC. In addition, an external reviewer with expertise in quantitative biology (QB) reviewed the assessment and provided recommendations to improve assessment items. QM BUGS III was designed to provide data on growth of QB ability, as well as identifying quantitative barriers that students encountered in developing QB ability (Mayes et al., 2020). This investigation provides evidence concerning the assessments' reliability and validity across the variables measured within the assessment.

QM BUGS III was deployed to investigate:

Research Question 1 (RQ1): What is the current state of students' ability to apply modelling in undergraduate biology?

Research Question 2 (RQ2): What is the current state of students' confidence in modelling in undergraduate biology?

Research Question 3 (RQ3): What is the evidence of reliability and validity for the QM BUGS III assessment?

Demographic variables and QM variables were used as comparison groups to support analyses addressing aspects of the three research questions. Comparisons using these variables included (1) how class standing (Freshmen, Sophomore, Junior or Senior) affected performance on the assessment tool; (2) gender, class standing, and race categories resulted in different performance or confidence levels; (3) whether there were associations among the variables of the assessment; (4) which quantitative skills were most difficult for biology students, and (5) whether validity evidence supported inferences for each variable assessed using the instrument.

Table 1. Population demographics of sampled students completing the QM BUGS III assessment.

Term	*University*	*Class standing (year)*	*Course level*	*Gender*	*Race*
Fl 2018 172	UNL 565	Yr 1 263	100 542	Male 222	White 490
Sp 2019 440	GSU 47	Yr 2 188	200 37	Female 386	Black 28
		Yr 3 93	300 0	NA 4	Hispanic 34
		Yr 4 68	400 33		Native American 2
					Asian 35
					Other 14
					NA 9

Class standing, gender, and race were self-reported.
UNL: University of Nebraska-Lincoln; GSU: Georgia Southern University.

Methods

Sample

Across one academic year 612 students from two universities consented to participate in the research (**Table 1**). The largest percentage of students self-reported as freshmen (43%) or sophomore (31%). The participants were majority female (64%) and white (80%).

Students were sampled predominantly from biology courses at the first-year course level suggesting introductory materials meant to establish a base of knowledge for life science students. We used convenience course sampling with five instructors and some of these courses were at the sophomore and senior level representing more context specific courses like ecology and population dynamics. Students were enrolled at a large public Midwestern university (UNL) and a large public southeastern university (GSU) (**Table 2**).

Instrument

Scoring, measurement, and analyses

The QMBUGS III instrument was administered digitally on personal computers outside of class time, which could introduce bias since it was not in a controlled classroom setting. Examination of the raw data

Table 2. Number of participants at students' self-reported class standing and enrolled course level.

Class standing	*First year* *100 level*	*Second year* *200 level*	*Fourth year* *400 level*	*Total*
Freshmen	263	–	–	263
Sophomore	179	9	–	188
Junior	66	17	10	93
Senior	34	11	23	68
Total	542	37	33	612

indicated that missing data was relatively minimal, ranging from 0 to 7 non-responses per item. Normality of item data was examined using skewness, kurtosis and Shapiro–Wilk statistics. All items were non-normally distributed at the p < 0.001 level. For MP and MMK 18 items (72%) were skewed positively and 16 items (64%) had a negative kurtosis. For QBCC all 12 items (100%) were skewed negatively, all showing a positive kurtosis.

QA, QI and QM item sub sections of the assessment were each examined and compared, but were also combined to form the modelling practices (MP) variable, defined as reasoning about biological phenomena using quantitative accounts of relationships in models and modelling in undergraduate biology. MP items were analyzed as correct/incorrect with 20 as a perfect score.

The MMK subsection included five multiple-response items for which students received 0.2 credit for correctly choosing or not choosing each of the five possible responses. This resulted in 6 ordinal levels of performance for each item (0, 0.2, 0.4, 0.6, 0.8, 1.0). A sixth item where students arranged responses to indicate the major steps of the modelling process was interpreted by examining patterns in students' responses.

QBCC questions asked participants how capable they were in modelling. Responses reflected an overall self-rating of capability confidence to perform quantitative modelling in biology tasks. The 12 Likert items on QBCC were on a 4-level scale with 4 being the most positive response, yielding a perfect score of 48 across the items.

Unidimensional Rasch calibrations of the raw data were conducted using Winsteps software (Linacre, 2017) to enable construction of interval measures of item-difficulty and person-ability, and to calculate Rasch diagnostics that include reliability and separation indices, point-measure correlations, and item fit indices (RQ3) (Bond & Fox, 2015; Engelhard, 2013). The Rasch measures represent an interval scale calibrated from the raw data to identify relative difficulties among items and participant abilities (RQ1). In addition, the Rasch calibrated item and person measures permit common scale comparisons of the item and student locations on a variable map to examine targeting of the instrument with the sample measures. The Rasch dichotomous model (Rasch, 1960) was used for calibrations of MP items as appropriate to their correct/incorrect (dichotomous) scoring. The Rasch rating scale model (Andrich, 1978) was used for calibrations of both MMK (6 scored levels) and QBCC (4 rated levels) as appropriate to those variables' ordinal category scoring each with a consistent number of ordinal levels across respective item sets. In general, though assessment performances typically depend upon several types of participant abilities, we are interested in evaluating each of our dependent variables using a unidimensional measurement model to identify whether each distinct variable can function as a predominant single dimension. We treat each variable, MP, MMK, and QBCC as separate constructs respectively.

Following the primary calibration analyses on the three main constructs, student performance comparisons (RQ1) were also analyzed using a within-subjects ANOVA of the MP factor that includes QA, QM, and QI items, using Bonferroni comparisons to avoid Type I error inflation. Effect sizes using the mean differences were then calculated to establish comparable difference magnitudes based on standard deviation units for both the within-subjects' differences on raw score student performances and the corresponding within-subjects Rasch-calibrated item-difficulty differences. Analyses of variance (ANOVA) tests were conducted and evaluated at a conservative $p < 0.01$ to avoid Type I error inflation while examining the independent variables of self-reported gender, class standing, and race category relative to Rasch calibrated MP, MMK, and QBCC measures to explore any unpredicted effects of group membership on assessment results (RQ2 and RQ3). These separate ANOVA tests were conducted on each dependent variable due to the insufficient group N necessary for a valid multivariate test of all three variables together. To test ANOVA assumptions for each of the three analyses Levine's test for equality of variances was conducted, Q-Q plots were examined, and a non-parametric Kruskal–Wallis test was used for corroboration of ANOVA due to the nonnormal data distributions. Statistical analyses were conducted using IBM SPSS (IBM Corp, 2017). Correlation plots and an ANOVA table were generated using JASP (JASP Team, 2019). Rasch calibrations, analyses, and graphics were conducted with Winsteps (Linacre, 2017). Effect sizes were calculated using the effect size calculator by Lenhard and Lenhard (2016).

Results

Difficulty measures

A variable map visually displays the item and person measures together to examine the ordered structure of items and to better understand the construct with respect to the people being assessed. Ideal targeting is nearly mirror-image distributions of items and students, including means and standard deviations, along the Rasch scale. The Rasch variable map of MP items and student measures (RQ1) shows the student distribution is located lower than the item distribution, indicating that collectively the items were difficult for students (**Figure 1**). QA, QI and QM all included items that were distributed approximately 1.5 logits, with the QA distribution slightly lower than that of QI and QM. There were five items that were at or above one standard deviation above the mean (2 QI, 3 QM). The four easiest items were at or below one standard deviation below the mean (2 QA, 1 QI, 1 QM).

Rasch variable maps of MMK items and QBCC items (Appendix A) both illustrate how respective student distributions were located above the corresponding item distribution on each map's logit scale. The MMK distributions (RQ1) showed how most students were aware of the nature and utility of models, the qualities required for a model to be acceptable, and the characteristics of a quality model. They were less aware of model components, purpose, and characteristics. The QBCC distributions (RQ2) showed how most students were

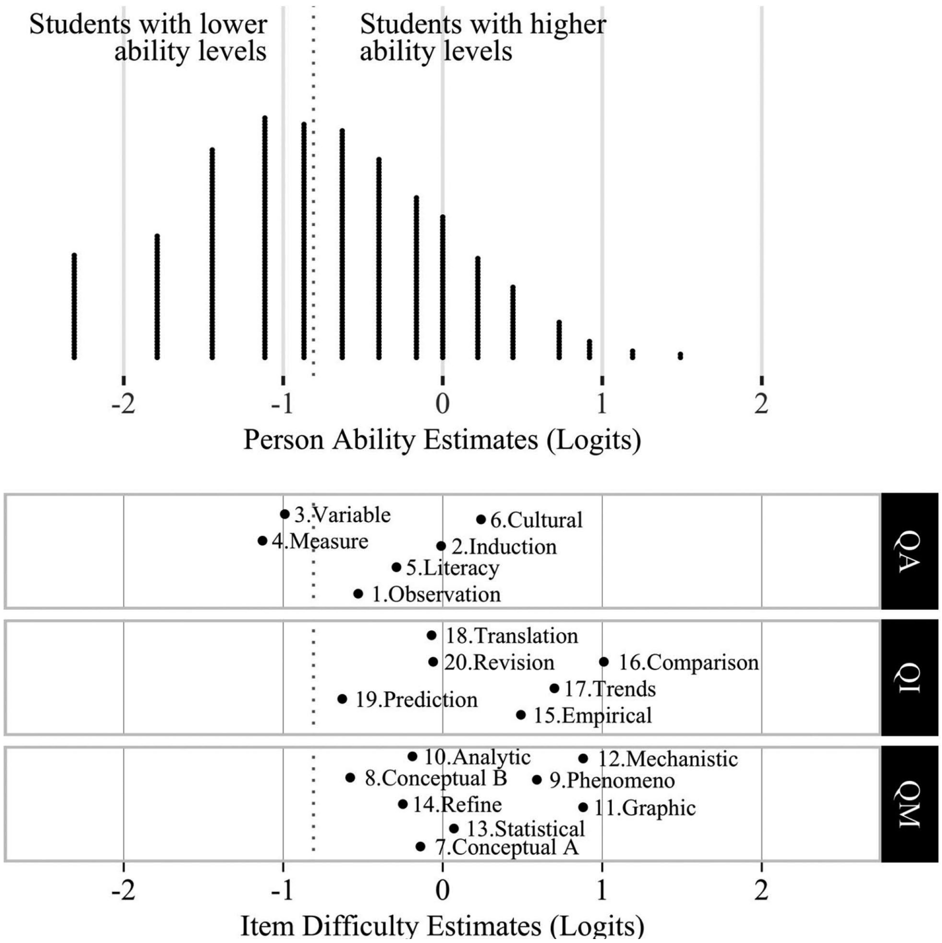

Figure 1. Variable map of modelling practices (MP) illustrating the logit distribution of student ability measures (top) and item difficulty measures (bottom), with highest ability and difficulty on the right. Items (including numbering and brief descriptor) are grouped by Quantitative Act (QA), Quantitative Interpretation (QI), and Quantitative Modeling (QM). Overlap of these two distributions (targeting) supports measurement accuracy and reliability to the greatest degree at higher abilities and difficulties. The vertical dotted line indicates the mean ability level and highlights that all but two items were relatively difficult.

confident in their capabilities with quantitative biology, being most confident in creating a model, refining an existing model, and determining trends from a model, and less confident in their ability to use descriptive statistics to describe a data set or in developing a testable hypothesis.

Table 3. Rasch reliability and separation for three sections of the QM BUGS III assessment showing stronger reliability and separation for items than persons.

	MP	*MMK*	*QBCC*
Person Reliability	0.56	0.41	0.81
Person Separation	1.12	0.84	2.04
Item Reliability	0.97	0.99	0.99
Item Separation	6.13	8.32	8.72

MP: modelling practices and includes quantitative act, quantitative interpretation, and quantitative modelling; MMK: Metamodelling knowledge; QBCC: Quantitative biology capability confidence.

Reliability of measures (RQ3)

Person and item reliability and separation indices were calculated to examine reproducibility of the assessment's measures (**Table 3**). The person indices reflected whether the test discriminated the student sample into a sufficient number of levels for the test's purpose. The item indices reflected whether the student sample size was sufficient to precisely locate the item measures. Reliability indices range from 0 to 1, with levels of 0.8 and above being considered ideal (Bond & Fox, 2015; Linacre, 2017). Separation indices reflect the actual number of levels discriminated and range from 0 and above, with 1.5 representing an acceptable level, 2.0 representing a good level (i.e. two groups discriminated), and 3.0 or above representing excellent levels (Boone et al., 2013). For QBCC items all four indices strongly supported reliability, while for MP and MMK items, only the two item indices were at strongly supportive levels above 3.0 for item separation and above 0.8 for item reliability (Table 3). Person separation and reliability levels were below acceptable levels for MP and MMK. Thus, the performance variables were not able to consistently discriminate this student sample ideally into distinguishable groups, despite the precision identified in locating the item measures. However, the self-rating items were able to both distinguish two separate groupings of QBCC level, students with a lower level of confidence and those with a higher level of confidence, and locate levels precisely on the measurement scale. In addition, the strong item reliability and separation findings across each section support the item difficulty hierarchies reflected in the variable maps.

Item point-measure correlation (RQ3)

Point-measure correlations were used to identify the degree to which each item functioned in alignment with the respective instrument section, helping to distinguish each modelling abilities variable and the capability confidence variable. Positive correlations indicate favorable item functioning with higher levels preferable, ideally at 0.50 or above, to those nearly zero or negative. Items with correlations less than 0.15 were considered likely to be problematic and indicate need for item revision, replacement, or removal. All items across the instrument were correlated positively, with item correlations ranging from 0.10 to 0.50 for MP, 0.46 to 0.66 for MMK, and 0.59 to 0.73 for QBCC. Though MP item correlations were consistently positive, overall, they were weaker than those of the other two variables, with 4 of the 20 MP items near or below the 0.15 level, indicating items of possible concern.

Item fi*t*

How well each item fits the measurement model was examined as another means of evaluating each item's quality with respect to its contribution to measuring the respective variable. Fit was analyzed using the information weighted infit index and the outlier sensitive outfit index, each of which were reviewed using mean-square (MnSq) values. For MP and MMK items a MnSq range of 0.7–1.3 was used for both infit and outfit to assist the detection of problem items, and for QBCC a MnSq range of 0.6–1.4 was used per recommended guidelines by item type (Bond & Fox, 2015; Linacre, 2002). MnSq values of 1 are considered ideal fit for both types of items. The infit and outfit ranges were used to consider their relative fit alongside other measurement data. In addition, for this investigation we considered underfitting items (e.g. those above MnSq = 1.3) to be more potentially concerning than overfitting items, as underfit indicates items are not contributing to measurement of the variable.

The output tables in Appendix B include outfit statistics that reflect the item specific findings identifying misfitting items. We highlight outfit because of outfit's sensitivity to outliers (Linacre, 2002). Outfit mean, standard deviation, and maximum MnSq values for each variable summarize aggregate fit findings though the item specific statistics are crucial to these diagnoses. For MP items outfit MnSq = 1.03 (SD = 0.14), maximum = 1.34, with items 11 (MnSq = 1.31) and 12 (MnSq = 1.34) slightly over the MnSq = 1.3 threshold. For MMK items MnSq = 1.00 (SD = .16), maximum = 1.25, so no items were above the underfit threshold. For QBCC self-rating MnSq = 0.98 (SD = .17), maximum = 1.27, so no items were above the underfit threshold. For all three variables, the mean values were near to the ideal value of 1 and maximum values were within or just outside the expected range. In addition, no items were overfit by the Rasch model. These findings, in conjunction with the positive point-measure correlation and the person reliability findings are supportive of the measurement model for these three variables represented by QMBUGS III.

Differential item functioning

An examination of differential item functioning (DIF) was utilized to gain a particular awareness of the invariance of measurement, or how constant the item difficulty measures were across specific subgroups of students (females and males) within the student sample. Gender DIF (RQ3) was examined to identify any lack of invariance that might be a potential source of gender bias. The contrast represents a difference regarding how each respective item functions relative to the group of interest, holding constant the abilities between the two groups. Contrast values above 0.43 are considered low DIF, with values above 0.64 considered moderate to severe levels of DIF (Zwick et al., 1999). DIF statistical significance was examined using the Rasch-Welch t-test procedure (Linacre & Wright, 1989), a reliable measure for group sizes less than N = 300 (Schulz, 1990). DIF analyses were limited to the gender variable because of the relative similarity of the group sizes for males and females, in contrast to the more imbalanced group sizes for the race and class standing categories which would likely degrade reliability and the value of such findings.

For MP 19 items (95%) were free of gender differential item functioning (DIF). Item 9 had a gender DIF contrast value of 0.53 (p = 0.011) favoring performance by female students. MMK items were free of gender DIF with no contrast over a level of 0.06, strongly supporting measurement invariance. DIF analyses indicated that 10 QBCC

(RQ3) items (83%) were free of gender DIF. Item 35 had a small DIF contrast value of 0.47 (p < 0.01) that favored males and item 37 had a large gender DIF contrast value of 0.73 (p < 0.001) that favored females.

Dimensionality

To estimate each main variable as a dimension we used Principal Components Analysis (PCA) of residuals to determine the variance explained by measures and whether the first PCA contrast indicated a secondary dimension by an eigenvalue of 2 or greater, representing at least 2 items. In summary, the total variance explained for MP was 17.4% (eigenvalue of 4.19) with 6.5% (eigenvalue 1.57) unexplained in the first contrast. For MMK, the total variance explained was 35.2%(eigenvalue of 2.71) with 18.7% (eigenvalue of 1.44) unexplained in the first contrast. The total variance explained for QBCC was 42.1% (eigenvalue of 8.72) with 8.8% (eigenvalue of 1.83) unexplained in the first contrast. These PCA findings indicate that with the current sample QBCC and MMK explained more variance as dimensions than MP, and that secondary dimensions were not indicated for each variable.

Performance on modelling practices items: QA, QI, QM (RQ1)

The subsets of QA, QI, and QM items were examined with respect to student raw score performance means and Rasch calibrated item difficulty means. Student performance means were examined and reported as percentages to facilitate interpretability. In addition, Cohen's d effect sizes were calculated for MP subset mean comparisons examined, with the common characterizations of d = 0.20 as a small effect size, d = 0.50 as a medium effect size, and d = 0.80 as a large effect size (Cohen, 1988). Within the subsections, students performed best on QA (M= 42.53, SD = 26.07) and more modestly on QI (M= 29.33, SD = 19.37) and QM (M= 30.78, SD = 18.14).

The average student performance on the QM BUGS III assessment was relatively low for the modelling practices (MP: QA, QI and QM combined) items at 34%. Students performed best on the QA section of the assessment, correctly answering about 43% of the questions.

Table 4. Within subjects ANOVA for student performance on the modelling practices: QA, QI, and QM.

	Sum of square	*df*	*Mean square*	*F*	*p*	ω^2
QM factor	64222.028	2	32111.014	98.97	<0.001	0.070
Residual	396475.793	1222	324.448			

Note. Type III sum of squares.

Their performances on the QI and QM subsections were lower, only correctly answering about 30% of the items. ANOVA was evaluated at the conservative p < 0.01 level to account for the 3 within-subjects effects involving QA, QM and QI subset performance levels. Using raw score means of student performance a within-subjects ANOVA (**Table 4**) indicated statistical significance and Bonferroni post-hoc comparisons found QA performance differed significantly from QM (delta = 11.8, SE = 1.1), effect size of $d = 0.52$, t(1) = 11.0, $p < 0.001$ and from QI (delta = 13.2, SE = 1.1), effect size of $d = 0.58$, $t(1) = 12.0$, $p < 0.001$, while QI and QM were not significantly different (delta = 1.5, SE = 0.9), effect size of $d = 0.08$, $t(1) = 1.6$, $p = 0.34$. Overall, QI and QM performance levels were both over one half of a standard deviation below that of QA.

To determine whether these raw score mean comparisons of performance by each subset were consistent with those of constructed item measures (RQ3) per subset, Rasch item difficulty estimates were examined with respect to QA, QI, and QM items to allow for a comparative examination of these item subsets according to linear measures calibrated across all 20 MP items. Cohen's *d* effect sizes were then calculated using logit difficulty measures for standardized comparisons. QA items (M= −0.45, SD = 0.54) were less difficult than QI items (M= 0.24, SD = 0.60), $d = 1.21$, QA items were less difficult than QM items (M= 0.16, SD = 0.56), $d = 1.11$, and QM items were similar though slightly less difficult than QI items, $d = 0.14$. These three effect sizes for comparisons of item difficulty measures were each relatively larger than the effect sizes between corresponding subset comparisons of person performance raw score means. In addition, the least difficult item subset, QA, corresponded with the highest levels of overall performance among QA, QI, and QM, as expected.

Gender, race category, and class standing

Analysis of variance (ANOVA) was used to examine for possible effects of gender, race category, and class standing for each of the dependent variables of MP, MMK, and QBCC (Appendix C). Analyses of interactions including gender x race category, gender x class standing, and race category x class standing were also examined. Gender, race category, and grade level analyses did not yield statistically significant differences or interactions with MP, MMK, and QBCC variables (Q1 and Q2). Levine tests were each non-significant supporting homogeneity of variances, and Kruskal–Wallis non-parametric tests corroborated the ANOVA findings for each dependent variable.

Performance on metamodelling (RQ1)

The overall mean percentage on MMK items was 60.6 (SD = 12). Item 26 in MMK asked students to display their knowledge of the modelling process by arranging seven key steps in the process in a typical order: (1) Formulate hypothesis, (2) Identify variables, (3) Run experiment, (4) Analyze data, (5) Create model, (6) Interpret findings, (7) Revise model (**Figure 2**). All of the 653 students who completed item 26 were included in the analysis. A qualitative analysis of student ordering was intended to provide insight into students' thinking of the modelling process. We acknowledge there may be multiple 'correct' orders, and this analysis is descriptive of observed patterns of student responses. Only 8 students (1%) had five key steps in the correct position in the arrangement, and only 117 (18%) had 3 or more in the correct position.

Few students (59, 9%) identified formulate hypothesis as the first step. The most popular choices were interpret findings (32%) and identify variables (21%). These students are exploring the context for clues to a question they have not yet asked. A surprising percentage of students (43%) delayed formulate hypothesis until step 5 in the process. The most selected second step was run experiment (42%), indicating a desire to rush into the experiment without proper preparation. Only 21% of students correctly selected formulate hypothesis and then identify variables as the first two steps. The majority of students placed create model mid or late in the modelling process (57% step 4,

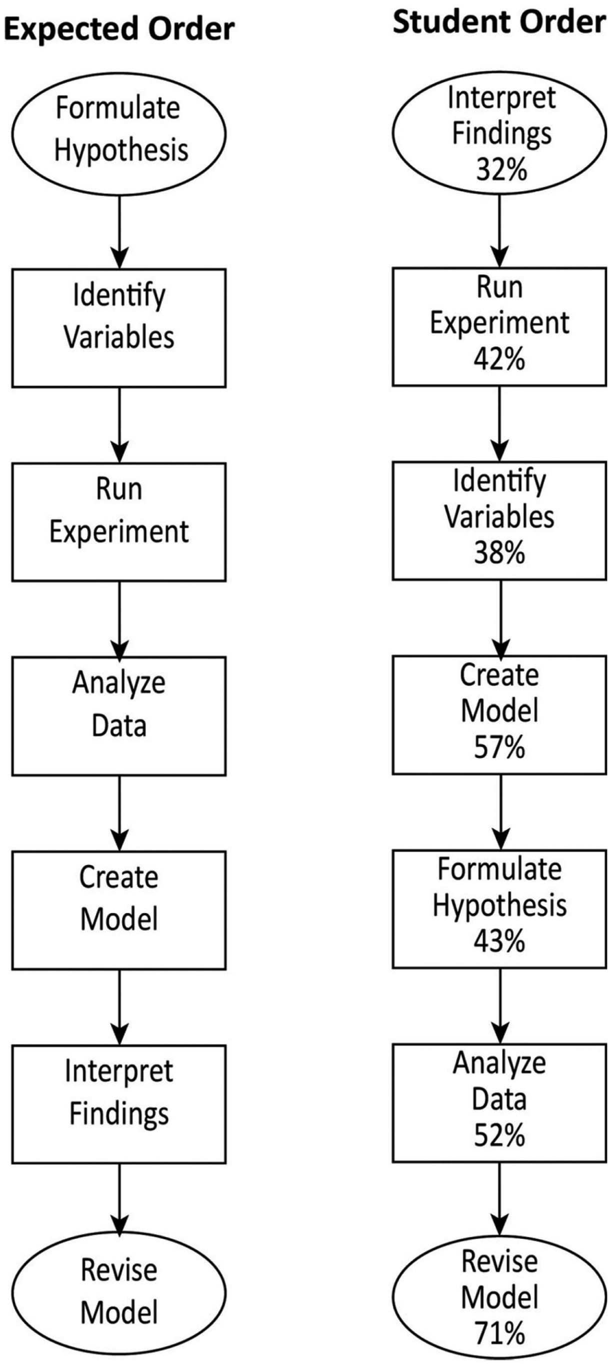

Figure 2. Modelling process network diagram including the expected order vs most frequently chosen student order. Student order (percent of students selecting given choice) based on most popular first choice, then most popular remaining choice for subsequent steps.

21% step 6). However, there were 11% who specified create model as the first step in the modelling process, possibly indicating they believe that models are to be provided by experts not developed by students, though other reasons for their choice are plausible as well. Of students who assigned create model as step 4, the steps preceding create model were: 81% run the experiment, 79% identify variables, 45% formulate hypothesis, and 44% interpret finding. The location of interpret findings is perplexing and raises questions about the purpose of the model for students. The most correctly placed step was revise model (71%), located in the final position. But the sixth step preceding revise model was expected to be interpret findings, and that was selected by only 17% of students. The most popular choice proceeding the final step was analyze data (52%). Perhaps students were confusing verifying the model, or worse plugging into the model, as analyzing data. The point in the modelling sequence at which one interprets findings was one of the most uncertain steps for students, being placed too early in the sequence by 83% of students. Perhaps students see the model as a product of the experimental findings rather than a tool to interpret the biological phenomena.

Levels of self evaluation of quantitative biology capability confidence (RQ2)

The overall mean percentage on the QBCC items was 75.5 (SD = 10.4). Students overwhelmingly agreed or strongly agreed with the 12 ‘I am capable… ’ statements that comprised QBCC (mean score 76%). On nine of the statements, greater than 85% of students responded agree or strongly agree, suggesting they felt capable of such actions as using descriptive statistics and statistical tests, identifying variables, making predictions from data, developing testable hypothesis, and translating between models. Among the high student confidence items was being capable of reasoning with models to improve understanding of the real-world, which indicates confidence in QI ability. The three lowest rated items all connected to students’ belief in their capabilities in creating a model. Only 72% of students responded Agree or Strongly Agree that they were capable of creating their own model (item 32; M= 2.81 out of 4), while 76% felt capable of refining a model to extend it to a new situation (item 34; M= 2.83). An interesting chasm arose between students feeling capable of determining trends and defending those trends, with over 85% indicating strong confidence in doing this when using biological arguments, and only 73% felt capable when using mathematical arguments (M= 2.83).

Associations between instrument sections (RQ1 and RQ2)

Pearson correlations were examined to determine the magnitude and direction of associations between the variables of the instrument. Using correlations, we examined whether and how much the variables increased and decreased together to further understand the relationships among the modelling abilities. There was a significant positive correlation among raw score means of all three MP subsections: QA and QM (r = 0.33, p < 0.001), QA and QI (r = 0.32, p < 0.001), and QM and QI (r = 0.28, p < 0.001). These correlation levels provide evidence of a consistent strength of relationships between pairs of MP subsections. There were also significant positive correlations between raw score means for each of QA, QI, and QM subsection items respectively with MMK section items (Appendix D), again with the QA correlation the strongest among the three associations. Using Rasch measures of the three main variables, MP was significantly correlated with MMK (r = 0.41, p < 0.001) but not with QBCC self-ratings (r = 0.015, p = 0.716). QBCC was not significantly correlated with MMK (r = −0.036, p = 0.376; **Figure 3**).

Discussion

Modelling practices (RQ1)

Student performance on the QA subsection was higher than that of QI or QM perhaps due to the call for training, explicitly or implicitly, in quantitative literacy throughout secondary school (National Council of Teachers of Mathematics, 2000). Despite this type of exposure students were still largely underprepared for QA (Hughes-Hallett, 2003; Steen, 2004). The percentage correct for QA is reflective of the 54% correct found by Johnson and Kaplan (2014) in a study of quantitative literacy among undergraduate statistics students. Performance on the

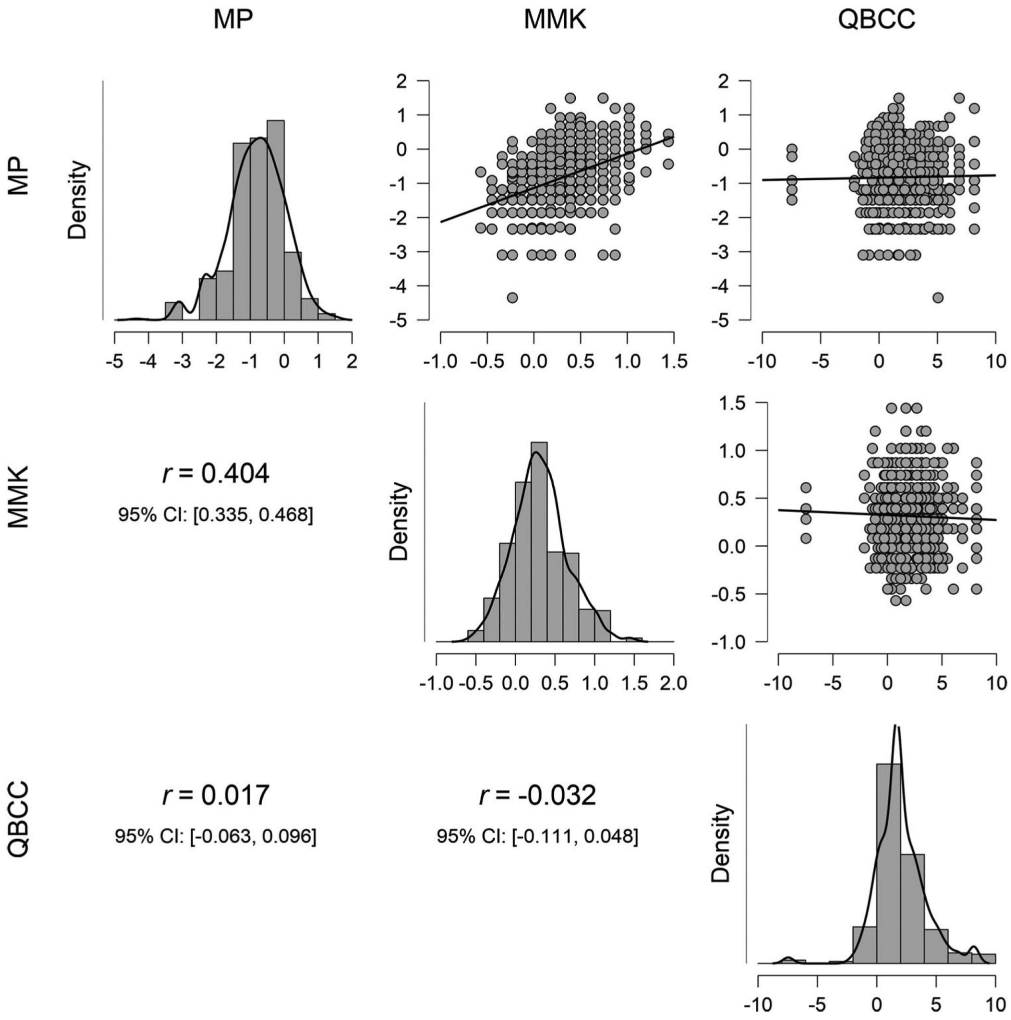

Figure 3. Pearson correlations (with confidence intervals, lower diagonal), and scatter plots with best fit line (x-axis = column title, y-axis = row title, upper diagonal) and distribution of raw scores (diagonal) among cumulative modelling practice (MP), metamodelling knowledge (MMK), and quantitative biology confidence capability (QBCC).

QA predicted performance on the QI and QM subsections, suggesting students who reasoned about quantitative relationships were better prepared to conduct interpretation and modelling (Appendix D) with the plant transpiration phenomena. Speth et al. (2010) found that incorporating quantitative literacy in undergraduate introductory biology courses through active-learning pedagogy improved quantitative skills, but construction of data-based scientific arguments was more of a challenge. Students in the present investigation performed particularly well on QA questions that included anchoring their understanding of transpiration by identifying a hypothesis, identifying relevant variables to study, and quantifying the variable of interest (Figure 1). The students were more comfortable with QA, given four of the five easiest items for students were located on the measurement continuum. However, even the two easiest items on quantifying a variable by determining an appropriate measure or identifying variables with attributes in context, were correctly answered by only 57% and 54% of students. In addition, it is worth considering that the superior QA performance may possibly be influenced by an order effect of QA being assessed first on the QM BUGS III assessment.

Given the push in biology to consider quantitative reasoning in the form of graphical analysis (AAAS, 2011; National Research Council (NRC), 2003) including how to interpret functional relationships (Stanhope et al., 2017), and results from deployment of QM BUGS I and II, we expected that students would perform at an intermediate level on the QI subsection. Stanhope et al. (2017) found that items related to visualizing data generally had low difficulty although a few questions about translating between a research question to a visual model had a high difficulty. However, this intermediate level performance was not found in the present study and that likely contributed to the poor performance on QM. QM is not considered entirely dependent on QI performance in a hierarchical manner, although there are elements of QI that are important for QM elements (Mayes et al., 2015). For example, the ability to refine a model will be predicated partially on one's ability to interpret the trends of the current model and make a mental prediction about a new model or new data. Goldstein and Flynn (2011) found that even students who learned quantitative analysis skills had trouble applying the skills to interpret biological datasets. At the same time, refining a model also involves additional cognitive tasks like knowing the nature of models (quantitative relationships, biological significance) and the purpose of determining coherence with scientific evidence.

There were bright spots in the QI and QM subsections as students performed well on the QM item related to a conceptual model of transpiration and the QI item related to extrapolation beyond the available data set (Figure 1). However, the QI and QM sections contained the five most difficult items. QM item 11 (18% correctly responded) had a minimal degree of discrimination, as indicated by the low point-measure correlation of 0.10 and focused on identifying a graph model for a table of data. Students struggled with determining which graph reflected the trends in the data table, another indication that graphic representations are challenging for students (Picone et al., 2007). Students overwhelmingly selected the linear model distractors instead of fitting multiple possible models then deciding based on fit. QM item 12, on which only 18% responded correctly, assessed mechanistic model building, that is, creating a model from theory. Students were provided a set of three relationships and asked to select a model based on theoretical constraints, that is to develop a mechanistic model based on first principles. The difficulty of this item was not surprising given the challenges of developing a mechanistic model.

QI item 16 was the most difficult for students at 1.01 logits, with only 17% of students responding correctly to it, and a low discrimination level of 0.12. This item assessed the QI ability of making model comparisons and engaged students in comparing equation models with a numeric table. They struggled with identifying the best evidence for fit between the two models. QI item 17, which 21% answered correctly, focused on ability to apply graphic and equation models to determine trends. Both included more than one model representation type and students selected all distractors on item 17 suggesting lack of understanding.

(RQ3) The variable map from the Rasch analysis of modelling practices indicated imperfect targeting between persons and items, with the item difficulty distribution located primarily on the high end of the person ability distribution. For this sample, the items were too predominantly difficult to support reliable ability measure estimates at the lower end of ability, suggesting the assessment may be better suited to higher ability students. This is a potential concern both about the instrument and student ability in QM. The QM BUGS III items were developed based on multiple frameworks and have been tested and revised multiple times (Mayes et al., 2019). The items reflect the expectations of QM experts for student outcomes. Yet students performed at a low level on the items. Students with junior and senior standing performed similarly to freshmen and sophomore students, indicating that QM ability is not improving due to exposure in a biology program. This is an area in need of further investigation including how students develop modelling and quantitative modelling ability as they progress through a biology curriculum.

Metamodelling knowledge (RQ1)

Students performed comparatively well on the MMK subsection, suggesting an awareness of the nature and purpose of modelling (Schwarz & White, 2005). The mean score on MMK items was 60.6%, indicating significantly better student performance then on MP items. Most students (greater than 70%) recognized models contain concepts and have a representation like an equation, diagram, or graph. Students were split on whether experiments are characteristics of models. Nearly 75% of students recorded that models consisted of objects and processes among objects, but far fewer (42%) selected that models included theories like the cohesion of water that would govern the processes. MMK item 23 was the most difficult for students at 0.32 logits (49% of students responded correctly), asking students to identify equation model acceptability. Approximately half of students recognized that equation models do not match the data collected, do not need to include all variables explaining the phenomena, explain research observations, are consistent with theories and other models, and are predictive. Item 24 had similar results and was more generalizable to all models, although the percentage who correctly selected each option was much higher than it was for item 23. Students indicated they believed that models were ideas explaining phenomena, not just equations or graphs. When asked about the purpose and utility of a model, approximately half of students (56%) agreed that models influence and constrain future research and 64% found models help explain reality. Nearly 70% selected that models were based on collected data with the purpose of predicting future events and 59% selected that models were assessed on their ability to explain real-world phenomena.

The variable map indicated good but imperfect targeting of persons to items, with the item difficulty distribution located primarily on the middle to lower end of the person ability distribution, but not on the high end of the person abilities (Appendix A). Thus, for this sample, the items were too predominantly easy to support highly reliable ability measure estimates at the upper end of ability. However, there was relatively favorable targeting for those at the middle and lower levels of ability. The instrument represents a broad dimension (e.g. metamodelling knowledge) reasonably well, with fit and strong item correlation providing additional support for the MMK variable. Instrument fairness and lack of bias was supported by invariant item functioning relative to male and female performance. Given these favorable targeting, dimensionality, and invariance findings, the graduated credit scoring (rather than dichotomous, correct/ incorrect) approach used with metamodelling items appears to have been advantageous for measurement in support of inferences from data.

Metamodelling knowledge correlated strongly with students' MP score (r = 0.42).When students were aware of how modelling impacts science knowledge advances, they were more likely to implement it (Schwarz & White, 2005). Schwarz and White (2005) advised explicitly addressing metamodelling knowledge to improve students' understanding of the overall culture of science. Fortus et al. (2016) found that given the proper support, MMK is attainable and improves the practice of modelling within the content area in which it is provided. This presents a teaching opportunity where instructors can frame biology instruction around the lens of modelling. While MMK alone is not sufficient, it does appear to be necessary as students perform quantitative modelling in biology.

Quantitative biology capability confidence (RQ2)

Clearly students hold their capabilities in high regard, but that did not match performance on the MP or MMK subsections as reflected by the lack of correlation strength with QBCC (Figure 3). The mismatch between ability and confidence for students is well established (Kennedy et al., 2002; Kruger & Dunning, 1999), and has been confirmed in biology (Chaplin, 2007) although others have found a correlation between math confidence and performance on a post-course quantitative skills assessment (Flanagan & Einarson, 2017). Even when students are confident in their abilities, they have to see value in work (Wigfield & Cambria, 2010) and our study did not directly measure their value towards using math in a biological context. The relatively high ratings on QBCC may reflect a lack of perceived alignment of the items requiring self-ratings of capability with specific topics as understood by these students. Students' estimates of their general capabilities may not have been influenced by their estimates of performance on specific questions. When students were asked about their own capabilities in an area, as they were on these items, they may have interpreted these questions with respect to general potential rather than how they performed on a related problem.

Implications

Quantitative modelling is a challenging endeavor requiring modelling abilities, biology knowledge, mathematical knowledge, quantitative reasoning ability, and confidence in working at the intersection of these. For biology students, it is clear that they have rarely mastered all the dimensions. Perhaps this should not be expected for students enrolled in lower division biology courses, despite having been exposed to quantitative skills in biology at the K-12 level. However, there are ample opportunities to support student development of quantitative modelling abilities at the collegiate level. Foremost, biology instructors can integrate modelling and quantitative reasoning in lower division courses. Students performed modestly on the QA items of the QM BUGS III assessment, suggesting this is an area of relative strength to leverage in learning QI and QM. Twenty-first century biology is an increasingly quantitative science, so undergraduate biology courses need to develop QA abilities to lay a foundation for students to interpret and build models. Speth et al. (2010) demonstrated that infusing existing course content and objectives with quantitative literacy concepts resulted in significant improvement in student ability to create graphical representations of biological data. Quantitative literacy is one component of QA, along with strengthening students' abilities in variable quantification, understanding covariation, and engagement in real-world contexts, which represent crucial opportunities to set the stage for modelling in biology.

Metamodelling knowledge is another area of relative strength and can be incorporated into biology curricula. Students who recognize the nature and purpose of the modelling endeavor have reason for greater motivation to rise to the challenge of quantitative modelling. Additionally, these students will tend to understand why they, as scientists, need to consider multiple dimensions of modelling like predicting, revising, interpreting, and creating models. We cannot delude ourselves into thinking that knowing steps of modelling will automatically improve modelling practices. Creating opportunities to practice many facets of quantitative modelling in biology courses throughout the curricula is the difficult work to be done. As Diaz Eaton et al. (2020) identified, there are professional development opportunities and resources to assist faculty wanting to engage students at the nexus of mathematics, quantitative reasoning, and biology. Clearly, the biology community needs to follow the trailblazing efforts of instructors who have used quantitative modelling as a way to engage students. The QM BUGS assessment can be a helpful tool to determine when those efforts support students' learning to quantitatively model biological phenomena.

* * * *

Acknowledgments We thank the instructors who graciously agreed to allow us to assess their students. We thank Autumn Fluent for her logistical and technical support. We thank A. Schuchardt and M. Aikens for providing expert reviews of the assessment. This work emerged from a working group organized by QUBES, and the attendees have directly and indirectly impacted this work throughout its development. Comments from two anonymous reviewers greatly improved this paper.

Disclosure statement No potential conflict of interest was reported by the author(s).

Data availability University of Nebraska-Lincoln Data Repository (QM BUGS assessment).

Supplemental data is attached to the archive record for this article.

References

AAAS. (2011). Vision and change in undergraduate biology education: A call to action. Washington, D.C. http://visionandchange.org/finalreport/

Acher, A., Arcà, M., & Sanmartí, N. (2007). Modeling as a teaching learning process for understanding materials: A case study in primary education. Science Education, 91(3), 398–418. https://doi.org/10.1002/sce.20196

Association of American Medical Colleges & Howard Hughes Medical Institute. (2009). Scientific foundations for future physicians: Report of the AAMC-HHMI committee. Washington, D.C. and Chevy Chase. Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43(4), 561–573. https://doi.org/10.1007/BF02293814

Bamberger, Y. M., & Davis, E. A. (2013). Middle-school science students' scientific modelling performances across content areas and within a learning progression. International Journal of Science Education, 35(2), 213–238. https://doi.org/10.1080/09500693.2011.624133

Bond, T., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences. Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). Rasch analysis in the human sciences. Springer.

Chaplin, S. (2007). A model of student success: Coaching students to develop critical thinking skills in introductory biology courses. International Journal for the Scholarship of Teaching and Learning, 1(2), Article 10. https://doi.org/10.20429/ijsotl.2007.010210

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Erlbaum.

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2016). Development of the statistical reasoning in biology concept inventory (SRBCI). CBE-Life Sciences Education, 15(1), ar5. https://doi.org/10.1187/cbe.15-06-0131

Diaz Eaton, C., Callender, H. L., Dahlquist, K. D., LaMar, M. D., Ledder, G., & Schugart, R. C. (2019). A "rule of five" framework for models and modeling to unify mathematicians and biologists and improve student learning. Problems, Resources, and Issues in Undergraduate Mathematical Sciences (PRIMUS), 29(8), 799–829. https://doi.org/10.1080/10511970.2018.1489318

Diaz Eaton, C., LaMar, M. D., & McCarthy, M. (2020). 21st century reform efforts in undergraduate quantitative biology education. Letters in Biomathematics, 7(1), 55–66.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). Taking science to school: Learning and teaching science in grades K-8. National Academies Press. http://www.nap.edu/catalog/11625.html

Engelhard, G. (2013). Invariant measurement: Using Rasch models in the social, behavioral, and health sciences. Routledge. https://books.google.com/books?id=NlqtMAEACAAJ

Flanagan, K. M., & Einarson, J. (2017). Gender, math confidence, and grit: Relationships with quantitative skills and performance in an undergraduate biology course. CBE—Life Sciences Education, 16(3), ar47. https://doi.org/10.1187/cbe.16-08-0253

Fortus, D., Shwartz, Y., & Rosenfeld, S. (2016). High school students' meta-modeling knowledge. Research in Science Education, 46(6), 787–810. https://doi.org/10.1007/s11165-015-9480-z

Fretz, E. B., Wu, H.-K., Zhang, B., Davis, E. A., Krajcik, J. S., & Soloway, E. (2002). An investigation of software scaffolds supporting modeling practices. Research in Science Education, 32(4), 567– 589. https://doi.org/10.1023/A:1022400817926

Garfunkel, S., & Montgomery, M. (2016). GAIMME report: Guidelines for assessment & instruction in mathematical modeling education. Consortium for Mathematics and Its Applications. http://www.siam.org/reports/gaimme.php

Gilbert, S. W. (1991). Model building and a definition of science. Journal of Research in Science Teaching, 28(1), 73–79. https://doi.org/10.1002/tea.3660280107

Gobert, J. D., & Buckley, B. C. (2000). Introduction to model-based teaching and learning in science education. International Journal of Science Education, 22(9), 891–894. https://doi.org/10.1080/095006900416839

Goldstein, J., & Flynn, D. F. B. (2011). Integrating active learning & quantitative skills into undergraduate Introductory biology curricula. The American Biology Teacher, 73(8), 454–461. https://doi.org/10.1525/abt.2011.73.8.6

Harrison, A. G., & Treagust, D. F. (2000). A typology of school science models. International Journal of Science Education, 22(9), 1011–1026. https://doi.org/10.1080/095006900416884

Hester, S., Nadler, M., Katcher, J., Elfring, L., Dykstra, E., Rezende, L., & Bolger, M. S. (2018). Authentic inquiry through modeling in biology (AIM-Bio): An introductory laboratory curriculum that increases undergraduates' scientific agency and skills. CBE - Life Sciences Education, 17(4), ar63. https://doi.org/10.1187/cbe.18-06-0090

Hoffman, K., Leupen, S., Dowell, K., Kephart, K., & Leips, J. (2016). Development and assessment of modules to integrate quantitative skills in introductory biology courses. CBE-Life Sciences Education, 15(2), ar14. https://doi.org/10.1187/cbe.15-09-0186

Hughes-Hallett, D. (2003). The role of mathematics courses in the development of quantitative literacy. In B. L. Madison & L. A. Steen (Eds.), Quantitative literacy: Why numeracy matters for schools and colleges (pp. 91–98). National Council on Education and the Disciplines.

IBM Corp. (2017). IBM SPSS statistics for windows (Version 25) [Computer software]. IBM Corp.

JASP Team. (2019). JASP (0.11) [Computer software].

Johnson, Y., & Kaplan, J. (2014). Assessing the quantitative literacy of students at a large public research university. http://www.statlit.org/pdf/2008JohnsonKaplanCRUME.pdf

Jonassen, D., Strobel, J., & Gottdenker, J. (2005). Model building for conceptual change. Interactive Learning Environment, 13(1–2), 15–37. https://doi.org/10.1080/10494820500173292

Kennedy, E. J., Lawton, L., & Plumlee, E. L. (2002). Blissful ignorance: The problem of unrecognized incompetence and academic performance. Journal of Marketing Education, 24(3), 243– 252. https://doi.org/10.1177/0273475302238047

King, G. P., Bergan-Roller, H., Galt, N., Helikar, T., & Dauer, J. (2019). Modelling activities integrating construction and simulation supported explanatory and evaluative reasoning. International Journal of Science Education, 41(13), 1764–1786. https://doi.org/10.1080/09500693.2019.1640914

Koponen, I. T. (2007). Models and modelling in physics education: A critical re-analysis of philosophical underpinnings and suggestions for revisions. Science & Education, 16(7), 751–773. https://doi.org/10.1007/s11191-006-9000-7

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Lehrer, R., & Schauble, L. (2006). Cultivating model-based reasoning in science education. In R. K. Sawyer (Ed.), Cambridge handbook of the learning sciences (pp. 371–388). Cambridge University Press. http://psycnet.apa.org/psycinfo/2006-07157-022

Lenhard, W., & Lenhard, A. (2016). Calculation of effect sizes. https://www.psychometrica.de/effect_size.html

Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M. I., Snoep, J., Hucka, M., Le Novere, N., & Laibe, C. (2010). Biomodels database: An enhanced, curated and annotated resource for published quantitative kinetic models. BMC Systems Biology, 4, 92.

Linacre, J. M. (2002). What do infit and outfit, mean square and standardized mean? Rasch Measurement Transactions, 16(2), 878.

Linacre, J. M. (2017). Winsteps (4.0.1) [Computer software]. Winsteps.com

Linacre, J. M., & Wright, B. D. (1989). The equivalence of Rasch PROX and Mantel-Haenszel. Transactions of Rasch Measurement SIG, 3-2, 1–3.

Louca, L., & Zacharia, Z. (2012). Modeling-based learning in science education: Cognitive, metacognitive, social, material and epistemological contributions. Educational Review, 64(4), 471– 492. https://doi.org/10.1080/00131911.2011.628748

Louca, L., & Zacharia, Z. (2019). Toward an epistemology of modeling-based learning in early science education. In A. Upmeier zu Belzen, D. Krüger, & J. van Driel (Eds.), Towards a competence- based view on models and modeling in science education (pp. 237–256). Springer International Publishing. https://doi.org/10.1007/978-3-030-30255-9_14

Louca, L., Zacharia, Z., & Constantinou, C. P. (2011). In quest of productive modeling-based learning discourse in elementary school science. Journal of Research in Science Teaching, 48(8), 919– 951. https://doi.org/10.1002/tea.20435

Mayes, R., Dauer, J. T., Rittschof, K. A., & Gallant, B. (2020). Quantitative modeling in biology for undergraduate students (QM BUGS) diagnostic assessment. UNL Data Repository. https://doi.org/10.32873/unl.dr.20201008

Mayes, R., Forrester, J., Christus, J., Peterson, F., Bonilla, R., & Yestness, N. (2014). Quantitative reasoning in environmental science: A learning progression.

International Journal of Science Education, 36(4), 635–658. https://doi.org/10.1080/09500693.2013.819534

Mayes, R., Rittschof, K., Dauer, J., & Gallant, B. (2019). Quantitative modelling biology undergraduate assessment. Letters in Biomathematics, 6(1), 1–27. https://doi.org/10.1080/23737867.2019.1653234

Mayes, R., Rittschof, K., Forrester, J., Christus, J., Watson, L., & Peterson, F. (2015). Quantitative reasoning in environmental science: Rasch measurement to support QR assessment. Numeracy, 8(2), Article 4. https://doi.org/10.5038/1936-4660.8.2.4

National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics.

National Research Council (NRC). (2003). Bio 2010: Transforming undergraduate education for future research biologists. National Academies Press.

NGSS Lead States. (2013). Next generation science standards: For states, by states. http://www.nextgenscience.org/lead-state-partners

Oh, P. S., & Oh, S. J. (2011). What teachers of science need to know about models: An overview. International Journal of Science Education, 33(8), 1109–1130. https://doi.org/10.1080/09500693.2010.502191

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. https://doi.org/10.1111/j.1365-2729.2006.00217.x

Papaevripidou, M., & Zacharia, Z. C. (2015). Examining how students' knowledge of the subject domain affects their process of modeling in a computer programming environment. Journal of Computers in Education, 2(3), 251–282. https://doi.org/10.1007/s40692-015-0034-1

Penner, D. E. (2000). Explaining systems: Investigating middle school students' understanding of emergent phenomena. Journal of Research in Science Teaching, 37(8), 784–806. https://doi.org/10.1002/1098-2736(200010)37:8<784::AID-TEA3>3.0.CO;2-E

Picone, C., Rhode, J., Hyatt, L., & Parshall, T. (2007). Assessing gains in undergraduate students' abilities to analyze graphical data. Teaching Issues and Experiments in Ecology, 5(July), 1–54.

Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. Journal of Research in Science Teaching, 48(5), 486–511. https://doi.org/10.1002/tea.20415

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests (expanded 1980). University of Chicago Press.

Schuchardt, A. M., & Schunn, C. D. (2016). Modeling scientific processes with mathematics equations enhances student qualitative conceptual understanding and quantitative problem solving. Science Education, 100(2), 290–320. https://doi.org/10.1002/sce.21198

Schulz, E. M. (1990). DIF detection: Rasch versus Mantel-Haenszel. Rasch Measurement Transactions, 4(2), 107.

Schwarz, C., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., Shwartz, Y., Hug, B., & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. Journal of Research in Science Teaching, 46(6), 632–654. https://doi.org/10.1002/tea.20311

Schwarz, C., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. Cognition and Instruction, 23(2), 165–205. https://doi.org/10.1207/s1532690xci2302_1

Seel, N. M. (2017). Model-based learning: A synthesis of theory and research. Educational Technology Research and Development, 65(4), 931–966. https://doi.org/10.1007/s11423-016-9507-9

Sensevy, G., Tiberghien, A., Santini, J., Laubé, S., & Griggs, P. (2008). An epistemological approach to modeling: Cases studies and implications for science teaching. Science Education, 92(3), 424– 446. https://doi.org/10.1002/sce.20268

Sins, P. H. M., 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. https://doi.org/10.1080/09500690500206408

Speth, E. B., Momsen, J. L., Moyerbrailean, G. A., Ebert-May, D., Long, T. M., Wyse, S., & Linton, D. (2010). 1, 2, 3, 4: Infusing quantitative literacy into introductory biology. CBE-Life Sciences Education, 9(3), 323–332. https://doi.org/10.1187/cbe.10-03-0033

Stanhope, L., Ziegler, L., Haque, T., Le, L., Vinces, M., Davis, G. K., Zieffler, A., Brodfuehrer, P., Preest, M., Belitsky, J. M., Umbanhowar, C., Overvoorde, P. J., & Nehm, R. (2017). Development of a biological science quantitative reasoning exam (BioSQuaRE). CBE—Life Sciences Education, 16(4), ar66. https://doi.org/10.1187/cbe.16-10-0301

Steen, L. A. (2004). Achieving quantitative literacy: An urgent challenge for higher education. Mathematical Association of America.

Svoboda, J., & Passmore, C. (2013). The strategies of modeling in biology education. Science & Education, 22(1), 119–142. https://doi.org/10.1007/s11191-011-9425-5

Taylor, R. T., Bishop, P. R., Lenhart, S., Gross, L. J., & Sturner, K. (2020). Development of the BioCalculus Assessment (BCA). CBE—Life Sciences Education, 19(1), ar6. https://doi.org/10.1187/cbe.18-10-0216

Thompson, P. W. (2011). Quantitative reasoning and mathematical modeling. In L. L. Hatfield, S. Chamberlain, & S. Belbase (Eds.), New perspectives and directions for collaborative research in mathematics education. WISDOMe monographs (Vol. 1, pp. 33–57). University of Wyoming Press. http://www.uwyo.edu/wisdome/_files/documents/qr_reasoningmathmodeling_thompson.pdf

Tsui, C.-Y., & Treagust, D. F. (2013). Introduction to multiple representations: Their importance in biology and biological education. In D. Treagust, & C.-Y. Tsui (Eds.), Multiple representations in biological education (Vol. 7, pp. 3–18). Springer.

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. Developmental Review, 30 (1), 1–35. https://doi.org/10.1016/j.dr.2009.12.001

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 (5), 941–967. https://doi.org/10.1002/sce.20259

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. Journal of Educational Measurement, 36(1), 1–28. https://doi.org/10.1111/j.1745-3984.1999.tb00543.x