European Journal of Science and Mathematics Education, 2023, **11**(4), 717-737
ISSN: 2301-251X (Online)

OPEN ACCESS

Research Article

Modelling-based pedagogy as a theme across science disciplines–Effects on scientific reasoning and content understanding

Kathy L. Malone [1,2*]
0000-0002-9286-4054

Anita Schuchardt [3]
0000-0001-6978-6484

[1] University of Hawai'i at Hilo, Hilo, HI, USA
[2] Nazarbayev University, Astana, KAZAKHSTAN
[3] University of Minnesota–Twin Cities, Minneapolis, MN, USA
[*] Corresponding author: klmalone60@gmail.com

Citation: Malone, K. L., & Schuchardt, A. (2023). Modelling-based pedagogy as a theme across science disciplines–Effects on scientific reasoning and content understanding. *European Journal of Science and Mathematics Education, 11*(4), 717-737. https://doi.org/10.30935/scimath/13516

ARTICLE INFO

Received: 21 Dec 2022

Accepted: 21 Jul 2023

ABSTRACT

Due to the increased use of scientific models and modelling in K-12 education, there is a need to uncover its effects on students over time. Prior research has shown that the use of scientific modelling in K-12 classes is associated with improved conceptual knowledge and problem-solving skills. However, few studies have explicitly tested the longitudinal benefits of using model-based instruction on students' scientific reasoning skills (SRS) and content knowledge. This paper studies the effects of the use of modelling-based pedagogy in a longitudinal comparative case study on students' SRS using hierarchical linear modeling. Our findings showed that initial exposure to modelling-based instruction increased scientific reasoning scores significantly. By the end of their first year of science instruction, the average high school freshman in our study achieved the scientific reasoning level of many undergraduate STEM majors. More importantly, students in the lowest quartile of scientific reasoning demonstrated increased scores over the three years of the modeling-based course sequence. In addition, reasoning scores in the modelling classes were a significant predictor of post-content knowledge in all subjects. Our results suggested that students should be exposed to model-based instruction early and consistently to achieve equity in science instruction.

Keywords: modelling-based instruction, science models, scientific reasoning, secondary school

INTRODUCTION

Models and scientific reasoning instead of memorization of facts have become central in many international standards (e.g., KMK, 2005; Ministry of Education, 2014; Ministry of Education in Taiwan, 2018; NGSS Lead States, 2013). This shift towards authentic science practices was precipitated by international testing that revealed that students across the world were experiencing difficulty in their ability to use scientific evidence (OECD, 2019). The use of modelling in secondary classrooms has shown positive results across not only numerous science disciplines but also grade levels and has included improvements in conceptual performance, scientific reasoning skills (SRS), conceptual understanding, and problem-solving (e.g., Chang, 2010; Gobet et al., 2011; Hestenes et al., 1992; Jenkins & Howard, 2019; Malone, 2008; Malone & Schuchardt, 2020; Passmore & Stewart, 2002). Most of these studies have focused on short interventions and do not consider the effects of long-term usage of modelling on students' knowledge and SRS. SRS are important since they have been shown to correlate with conceptual gains and problem-solving (e.g., Nieminen et al., 2012;

Thompson et al., 2018). Unfortunately, even with additional content knowledge SRS do not seem to improve after secondary school (e.g., Ding, 2018; Ding et al., 2016; Norris et al., 2003). Thus, students' SRS must be as strong as possible by the end of secondary school to help ensure success in college and everyday life.

In this paper, we report on a longitudinal study conducted to investigate the effect of a multi-year model-based program on secondary school students' science content knowledge and SRS over three years. Given the increased use of model-based instruction in secondary schools internationally, this quasi-experimental longitudinal study should be relevant to all international contexts as countries move towards greater implementation of authentic practices.

BACKGROUND OF THE STUDY

In the following section, we discuss the theoretical background of scientific reasoning, models, and modelling. We then discuss the conceptual framework that guided this study.

Scientific Reasoning Skills

Research has defined scientific reasoning in multiple ways from hypothetico-deductive reasoning focused on hypothesis testing (Lawson, 2004) to the coordination of theory and evidence (Kuhn & Dean, 2004) to mechanistic reasoning focused on the generation of causal mechanisms (Russ et al., 2009). Lehrer and Schauble (2006) suggested that past research falls into three categories based on how scientific reasoning is defined. They specified the three categories as science as theory, science as logic, and science as practice. They believed that science as theory regards experiments as a theoretical test that can lead to conceptual change. Science as logic focused on the use of control of variables during experimentation. Finally, science as practice suggested that theory and reasoning including the use of scientific discourse would support the development of representations to make the phenomena visible (Lehrer & Schauble, 2006).

Regardless of the definition, there are commonalities in the scientific reasoning sub-skills found to promote conceptual change including hypothesis testing, control of variables or experimentation skills, data reasoning, evidence evaluation, hypothetical-deductive reasoning, probabilistic reasoning (the use of deductive logic and probability to handle uncertainty), and proportional reasoning (e.g., Lawson et al., 2000; Zimmerman, 2000). This study focused on the perspective of science as practice. We hypothesized that changing classroom norms across multiple years so that students "do science" through modelling-based instruction (MI) would increase students' engagement with data and evidence resulting in improvements in students' ability to use scientific reasoning subskills.

Studies have found correlations between students' pre-scientific reasoning skills and their final conceptual understanding across grade levels and subjects (e.g., Coletta & Phillips, 2005; Coletta et al., 2007; Ding, 2014; Getahun, 2022; Lawson et al., 2007; Moore & Rubbo, 2012; Nieminen et al., 2012; Orosz et al., 2023; Thompson et al., 2018; Williams et al., 2021). While these studies determined that the higher their incoming scientific reasoning abilities the more likely students would demonstrate greater conceptual change, other studies have shown improvements in scientific reasoning within courses can be elusive (e.g., Ding et al., 2016; Moore & Rubbo, 2012). For example, Ding et al. (2016) in a study consisting of 1,637 Chinese undergraduate students found that there was no significant increase in scientific reasoning scores across multiple universities and majors. However, Ding (2018) determined that while Chinese students from 4th to 8th grades did not show increases in scientific reasoning their reasoning abilities did increase from 8th to 12th grades. However, from 12th-grade onward reasoning ability was again static during undergraduate years.

However, some studies have shown that SRS can increase when students at all levels are taught with methods that focus on critical thinking skills (e.g., Blumer & Beck, 2019; Hester et al., 2018; Yanto et al., 2019). These correlations between prescientific reasoning and post-content scores raise the intriguing possibility that if students can increase their SRS early in their scientific careers they might be better prepared to excel conceptually in later years. This study was designed to begin to answer this intriguing possibility.

Scientific Models and Scientific Modelling

Models are idealized representations of scientific phenomena that have multiple uses in scientific practice including description, explanation, communication, and prediction (e.g., Giere, 2004; Svoboda & Passmore,

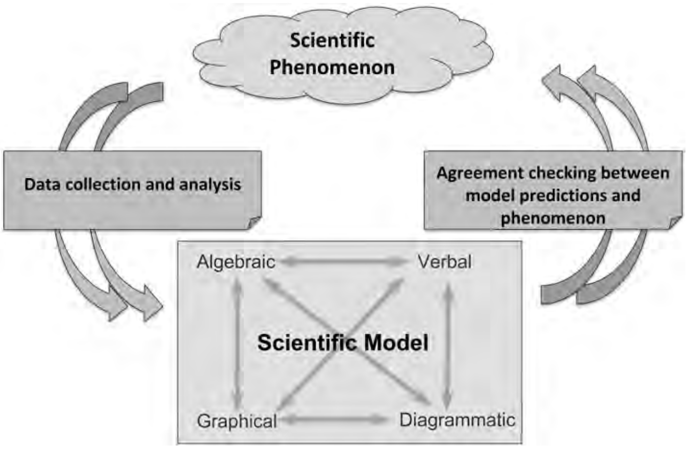

Figure 1. Modelling cycle (adapted from Malone et al., 2019)

2013). Traditionally, the term "model" has been used to refer to only physical representations such as a three-dimensional model of a cell. However, in scientific practice, a scientific model refers to a set of ideas about a scientific phenomenon. These ideas may be represented in multiple ways (i.e., physical models, mathematical equations, verbal descriptions, graphs, or pictures). No single representation offers a complete picture of the model, but together they provide a comprehensive portrayal that can be used to describe, explain, or make predictions (Buckley et al., 2004; Giere, 2004; Hestenes, 2010). Thus, DNA may be represented by a double helix physical model and a numerical ratio, or a written description (Giere, 2004; Godfrey-Smith, 2006). Experts fluidly switch between representations, understand that none is ever truly sufficient, and manipulate them to support their use as intellectual tools (Harrison & Treagust, 2000).

The term modelling refers to the iterative practices that scientists use to develop scientific models and their representations, known as the modelling cycle, illustrated in **Figure 1** (e.g., Clement & Steinberg, 2002; Halloun, 2007; Passmore et al., 2009). Scientists gather data about a real-world phenomenon and decide which entities, processes, or relationships to include in the scientific model. Multiple representations are developed as part of the modelling cycle. The representations developed during the modelling cycle for a scientific phenomenon have features in common (represented by double-headed arrows between representations) (Hestenes, 2010). Scientists can check the scientific model against observations and data produced through experiments (represented by the agreement arrows). The findings can be used to revise the scientific model and its representations or if the data fails to support the scientific model, a new scientific model can be developed (e.g., Halloun, 2007; Lehrer & Schauble, 2012; Passmore et al., 2009). We contend that the use of the modelling cycle depends on the deployment of and supports the development of SRS.

Connection Between Modelling-Based Instruction, Content & Scientific Reasoning–A Conceptual Framework

Students engagement in constructing models through participation in a modelling cycle has the potential to increase both students' learning of content and their ability to reason scientifically. During the iterative modelling cycle, students revisit the model content multiple times, thereby increasing the potential to confront alternative conceptions and master the content in greater depth (Halloun, 2007). Indeed, studies focused on the use of model-based curricula over a single course whether chemistry, biology, or physics have demonstrated increases in student content knowledge (e.g., Dori & Kaberman, 2012; Liang et al., 2012; Malone & Schuchardt, 2016).

Gains in SRS have not been studied as intensively as conceptual change in the case of MI. However, studies have shown that a student's scientific reasoning ability is a predictor of student performance when utilizing inquiry-oriented methods (e.g., Coletta et al., 2007; Ding, 2014; Moore & Rubbo, 2012).

During the development and testing of models, students are required to apply scientific reasoning sub-skills such as data analysis, construction, and evaluation of hypotheses, controlling variables, and deductive reasoning (Cameron et al., 2022; Dukerich, 2015; Malone, 2023; Posthuma-Adams, 2014; Svoboda & Passmore, 2013). Thus, having students engaged in constructing models during the modelling cycle has the

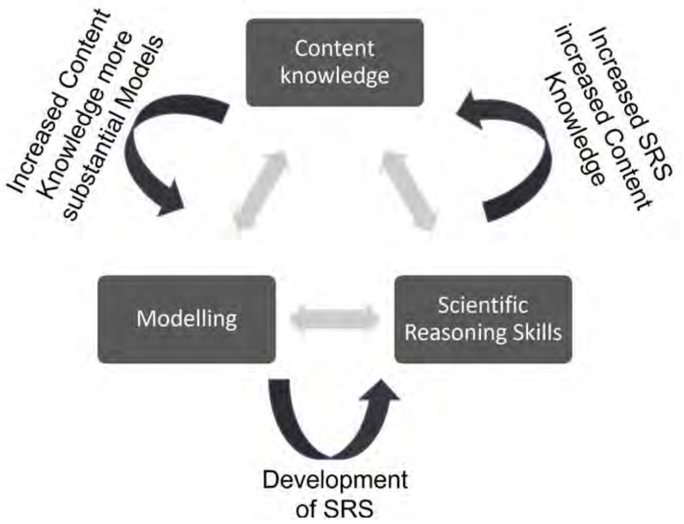

Figure 2. Conceptual framework (Source: Authors)

potential to increase not only students' learning of content but also their ability to reason scientifically. During the modelling cycle, students revisit the content supporting the model multiple times as well as deploy scientific reasoning sub-skills. For example, during the development and testing of models, (MI students apply SRS such as construction and evaluation of hypotheses, controlling variables, and hypothetico-deductive reasoning as well as data reasoning (e.g., Cameron et al., 2022; Dukerich, 2015; Malone, 2023; Posthuma-Adams, 2014; Russ & Odden, 2017; Svoboda & Passmore, 2013). Data reasoning includes proportional reasoning, pattern recognition, and probabilistic thinking. As the modelling cycle progresses from descriptive model representations to explanatory models' students develop a better understanding of when to revise models as well as how to use models to explain science content. In Schwarz et al. (2009), middle school students that followed a similar instructional progression improved their models and developed greater accuracy in predictions. They became more sophisticated in deploying the modelling cycle and started to be explicit about model limitations.

Given past research, there seems to be a cyclic connection between modelling, SRS, and content development. This cycle is the basis for this research study (see **Figure 2**).

Purpose of the Study

To date, there has not been any published longitudinal research investigating the shifts in students' SRS after participation in a sequence of classes grounded in modelling-based pedagogy compared to students who participated in non-modelling inquiry-based (NM-I) courses. Thus, a longitudinal quasi-experimental case study was developed to compare the shifts in scientific reasoning abilities as well as conceptual understanding of students taking an MI course sequence with those who completed an NM-I sequence from their first through the third year of high school science.

METHODS

This longitudinal quasi-experimental quantitative case comparison study used hierarchical linear modeling (HLM) statistical methods to compare two different three-year science course sequences. The research questions guiding this study were:

1. What were the effects on SRS of a modelling-based course sequence compared to a non-modelling (i.e., inquiry-based) course sequence?
2. What was the relationship between scientific reasoning ability and content knowledge in physics, chemistry, and biology modelling-based courses?
3. How did timing of the use of modelling affect SRS over the three years of the study?

Table 1. Teacher demographics

#	Modelling workshops attended	Education level	Years of high school teaching	Teaching assignment in inquiry cohort				Teaching assignment in modelling cohort		
				9th NM-I bio	10th NM-I chem	11th MI phys	11th NM-I phys	9th MI phys	10th MI chem	11th MI bio
1	1	PhD Bio	1-5	√				√		√
2	3	MA	10-20			√		√		
3	2	BS	5-10	√				√		√
4	1	MA	1-5	√	√			√	√	
5	1	MA	>20	√				√		√
6	0	MA	>20				√			
7	1	BA	10-20		√				√	
8	1	BA	5-10		√				√	

Participants and School Context

This study took place in an independent high school in the northeastern United States. 30.0% of the students received financial assistance and 27.0% were students of color. The study follows the school's transition from inquiry-based to modelling-based science instruction. Because models in physics are often more concrete than those in biology, the instructional sequence was shifted from a biology-chemistry-physics sequence to a physics-chemistry-biology sequence. This study follows the last cohort to take NM-I course sequence (n=119) and the first two cohorts enrolled in MI course sequence (n=252). The population was homogeneous in terms of socio-economic groups, ethnicity, and educational background. Each year 60.0% of the students came from a primary feeder middle school, while 40.0% came from other middle schools. In addition, students had either previously taken or were concurrently enrolled in algebra I. The schoolteachers, administrators, and researchers jointly selected the data collection instruments and the school included them as program assessments for both cohorts. Thus, the data analyzed in this study was derived from classroom artifacts (i.e., secondary data). The secondary data includes the use of test instruments chosen for each course such as released SAT biology questions, force concept inventory (FCI), and chemistry concept inventory (CCI) as well as grade-level exams taken by all students (such as independent school entrance exam (ISEE) and eighth-grade Educational Records Bureau's (ERB) comprehensive testing program (CTP). These exam artifacts will be discussed further in the methods section (BioSatII; Hestenes et al., 1992; Mulford & Robinson, 2002). Parental permission was given to analyze student data for publication purposes.

During the study, there were no changes made in other disciplinary courses. The history and English classes were taught using curricula that focused on analytical writing. The math courses were traditional teacher-centered courses without significant hands-on work.

Course sequences experienced by students

In both sequences, one author observed each teacher for an entire period at least three times a year. In addition, all course curriculum maps were available for the project staff to consult. In both sequences, there was no science teacher turnover during the years of the study.

Non-modelling inquiry course sequence: In NM-I instructional sequence, students' first high school science class was inquiry-oriented biology), followed by a more structured chemistry class during their second year. After the second year, some students opted not to enroll in a third course, but some did not. Those that did enroll in a science course during the third year normally enrolled in one of three options: biology II, chemistry II, or physics I. For physics I, the two physics teachers used different pedagogy. Teacher 2 used a modelling-based curriculum while teacher 6 used a non-modelling inquiry-oriented approach (see **Table 1**). Each of the courses in NM-I sequence were separate entities without any integration. See **Figure 3** for the timeline of these courses.

Modelling-based instruction course sequence: In MI sequence, the first course students took was modelling-based physics. The teachers who taught ninth-grade biology in NM-I sequence were required to take a three-week modelling instruction, a specific modelling-based pedagogy, physics workshop the summer before teaching ninth-grade modelling physics (Barlow et al., 2014; Hestenes et al., 1992). The chemistry

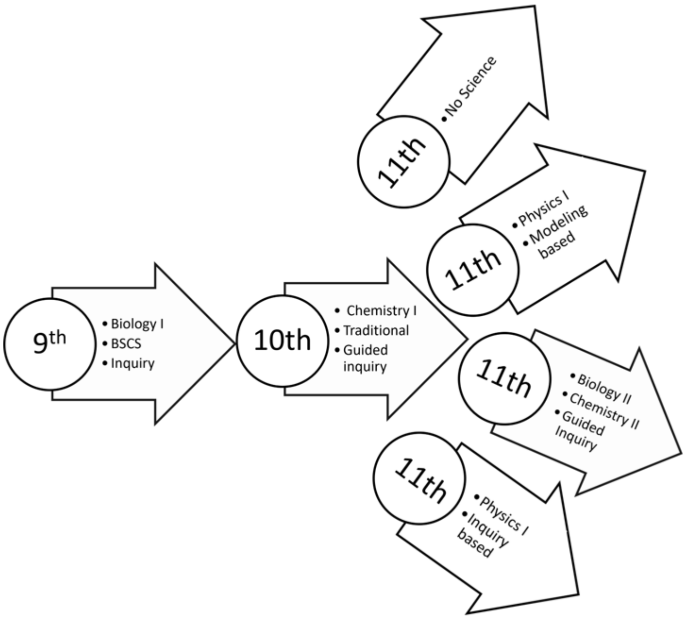

Figure 3. NM-I course sequence over the four years of the study (Source: Authors)

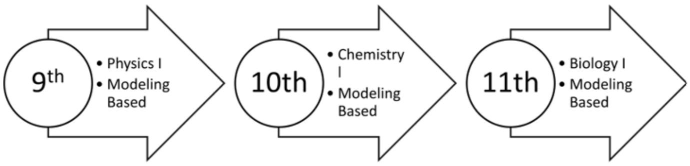

Figure 4. MI course sequence over the four years of the study (Source: Authors)

teachers who taught 10th grade chemistry in NM-I sequence were required to take a three-week Modelling Instruction chemistry workshop before teaching MI chemistry.

Because no MI biology course using modelling instruction pedagogy existed, teachers at the school who had been trained and previously taught the first-year MI physics class developed a third-year biology course grounded in the MI pedagogy. MI students took all three courses.

See **Figure 4** for the timeline of MI courses.

Comparison between the two sequences: The two sequences, NM-I and MI, had several similarities and differences in instructional strategies. NM-I sequence was a mix of pedagogies that is akin to what most students currently encounter during their secondary careers. On the other hand, MI sequence was integrated across years by the use of one pedagogical stance, which allowed students to become enculturated into the practice of scientific modelling. **Table 2** compares the classroom strategies utilized by both sequences based on inquiry practices highlighted by Furtak et al. (2012), Minner et al. (2010), and NRC (2015).

Instruments

Entrance exams

To compare the abilities of the entering freshmen, the mean performance on ISEE and eighth-grade ERB CTP assessments were compared. ERB CTP assessments were administered to all eighth-graders in the feeder middle school as part of ongoing assessment efforts. These scores represented the testing results of approximately two-thirds (60.0%-67.0%) of the incoming freshmen class each year. ISEE assessment was administered to all eighth-graders not enrolled at the feeder middle school (about 37.0%-40.0%).

Scientific reasoning ability

The investigators established several criteria that the assessment chosen for the study must meet. These criteria were, as follows:

Table 2. Comparison of course strategies in modelling & non-modelling inquiry courses (Y=yes, N=no, & R=rarely)

Strategy	Inquiry class sequence	Modelling class sequence
Instructional strategies		
Students watch scientific phenomena[2]	Y	N
Student-to-student talk dominates class discussions	N	Y
Teacher-to-student talk dominates the class discussions	Y	N
Emphasize student responsibility for learning[2]	N	Y
Drawing on/connecting to prior knowledge[1]	Y	Y
Eliciting students' ideas/mental models[1]	N	Y
Providing conceptually oriented feedback[1]	Y	Y
Participating in class discussions[1]	Y	Y
Arguing/debating scientific ideas based on evidence	N	Y
Presentations [1]	Y	Y
Working collaboratively[1]	Y	Y
Call and response formative assessment[2]	Y	N
Rote memorization of facts and terminology[3]	Y	N
Teacher provides information to whole class[3]	Y	N
Experimental strategies		
Asking scientifically oriented questions[1]	Y	Y
Experimental design[1]	Y	Y
Explicit about nature of science	N	N
Implicit about nature of science	N	Y
Drawing conclusions based on evidence[1]	R	Y
Students discuss each other's evidence	R	Y
Executing scientific procedures[1]	Y	Y
Recording data[1]	Y	Y
Representing data[1]	Y	Y
Students choose how to represent data	R	Y
Hands-on[1] and manipulate mateirals[2]	Y	Y

1. The assessment had to be domain-general and thus have items from chemistry, biology, and physics since it was to be used over the course of three years.
2. The assessment should include items targeting SRS such as hypothesis testing, control of variables, data reasoning, hypothetico-deductive reasoning, proportional probabilistic and correlation reasoning, model development, limitations, and affordances. It was realized that the ability to assess all these skills would be difficult to find within a single assessment.
3. The ability to compare the scores in this study to results from other studies from secondary through undergraduate college.
4. The assessment had to be completed within a class period of 45 minutes.

After reviewing several scientific reasoning assessments that have been previously developed and tested for reliability it was determined that Lawson's classroom test of scientific reasoning (LCTSR) was the best choice (Lawson et al., 2000). Reasons for this choice included that it had been used at multiple levels (e.g., Al-Balushi et al., 2017; Bernard & Dudek-Różycki, 2019; Ding et al., 2016; Illes et al., 2019), covered many of the SRS targeted by the modelling-based pedagogy (Opitz et al., 2017) and has been shown to be good for assessing scientific reasoning on a unidimensional scale (Bao et al., 2018; Ding, 2018). **Table 3** identifies the SRS of interest that are covered by LCTSR as well as those skills that are not associated with this instrument.

LCTSR was administered to MI students during the first weeks of their ninth-grade year and at the end of each course in the modelling sequence. It was administered to NM-I students at the end of each science course in all three years of the study. The lack of LCTSR results at the beginning of the 9th grade year for NM-I was due to a shift in high school administration policy that advanced the implementation of the physics first modelling sequence by one full school year. This administrative decision was made midway through the NM-I school year thus precluding a pre-assessment in the 9th grade year.

Correct responses on the 24-item LCTSR were awarded one point while incorrect responses were awarded zero points. **Table 3** shows the number of items that correlate with each SRS covered by LCTSR as well as the number of questions that were included in the categories identified in the original publication. The total

Table 3. Skills associated with Lawson's classroom test of scientific reasoning

Skills hypothesized to be affected by modelling pedagogy	Covered by LCTSR	Original LCTSR category (Lawson, 2000)	Number of questions on LCTSR for each skill	Number of questions in original categories
Nature of science				
Conservation	√	√	4	4
Hypothesis testing	√		6	
Control of variables	√	√	6	6
Hypothetico-deductive reasoning	√	√	4	4
Probabilistic reasoning	√	√	4	4
Examination of complex evidence				
Model development				
Model limitations & affordances				
Proportional reasoning	√	√	4	4
Argumentation skills				
Data reasoning	√		4	
Evidence evaluation	√		6	
Correlation reasoning	√	√	2	2

number of questions that cover all SRS skills is greater than the 24 questions that comprise the LCTSR because one question can cover more than one skill. Prior studies have analyzed student gains in different subskills since LCTSR allows for the grouping of questions by SRS (e.g., Bao et al., 2009; Ding, 2018; Sapia et al., 2022). While the LCTSR has been tested and validated in many studies (e.g., Kaygisiz et al., 2017; Lawson et al., 2000), to assess the reliability of this instrument in this context, all the post-test LCTSR scores from both sequences (n=962) were used to calculate the internal consistency of the questions via KR-20 reliability coefficient. The calculated KR-20 reliability coefficient of 0.83 was an adequate value for use in making comparisons (Schinka et al., 2003).

Content knowledge assessments

Content knowledge was assessed at the end of each year-long course (physics, chemistry, and biology) for the second cohort to enter MI sequence. The second cohort was chosen so that teachers would have at least one year of practice teaching with MI before their students were assessed for conceptual content gains. For physics, FCI was used (Hestenes et al., 1992). For chemistry, CCI was chosen (Mulford & Robinson, 2002). For biology, students took a practice SAT II biology exam (BioSatII). The calculated Cronbach's alpha for each assessment was as follows: FCI (alpha=.76, number of items=30, number of students=240), CCI (alpha=.83, number of items=22, number of students=276), BioSatII (alpha=.94, number of items=50, number of students=280). FCI has been used in several physics studies and classes worldwide since its publication in 1992 (e.g., Bouzid et al., 2022; Carleschi et al., 2022). FCI has 30 questions (Hestenes et al., 1992). These questions focus on Newtonian physics concepts as well as kinematics. Each question has item distractors that are connected to common student alternative conceptions. Gain scores on FCI have been correlated with pedagogical practices (Hake, 1998).

CCI is a 22-question survey that covers the transformation of matter, energetics, and representations in chemistry (Mulford & Robinson, 2002). Each question's distractors are tied to students' alternative conceptions of chemistry. CCI has also been used in multiple contexts since its inception (e.g., Aakre, 2021; Georgiou & Sharma, 2020; Mehl, 2022).

BioSatII consisted of 50 multiple-choice questions. The questions were taken from questions for SAT II biology assessment released by College Board, a private not-for-profit organization in the United States (College Board, 2020). SAT II biology assessment is designed to assess secondary students' ability to recall, comprehend and apply major biology concepts. 50 selected questions focused on concepts covered in the biology I classes in both sequences.

Data Analysis Techniques

Means, normalized gains, and standard error of the mean (SEM) and Cohen's d effect sizes were calculated. ANOVA was used to compare means and normalized gains. All assumptions were met. HLM was used to examine longitudinal gains. Regression analyses were performed to analyze the relative contribution of

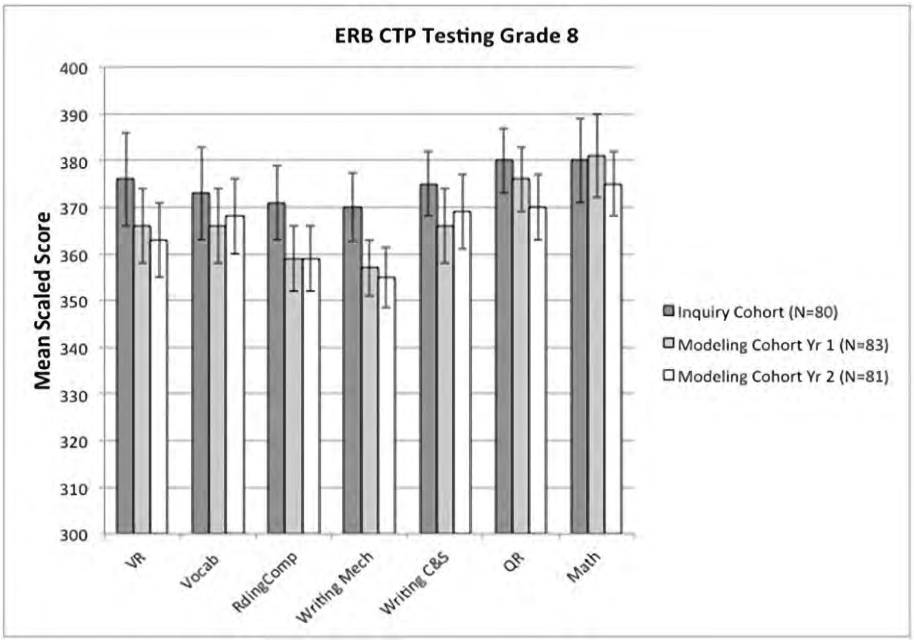

Figure 5. Mean ERB CTP grade 8 test scores for categories of verbal reasoning (VR), vocabulary (Vocab), reading comprehension (ReadingComp), writing mechanisms (Writing Mech), writing concepts & skills (Writing C&S), quantitative reasoning (QR), & mathematics (Math) (Source: Authors)

content knowledge and SRS on post-test performance. Detailed descriptions of each of these analyses are provided in the relevant results section to make the interpretation of the results easier.

HLM was used to answer the first research question concerning the effects of MI vs. NM-I course sequence on student SRS. Longitudinal HLM analysis was conducted on the scientific reasoning scores of all students with scores nested within students (Raudenbusch & Byrk, 2002) with exposure to modelling-based pedagogy, and timing of modelling-based pedagogy as independent variables, and time as a covariate. Students who only had one scientific reasoning score or students who did not have an initial test score were eliminated. All assumptions (e.g., homoscedasticity, normality, and independence) were met.

To answer the second research question concerning how the timing of the use of modelling affected SRS over the three years of the study we divided NM group into two subsections (i.e., those who took NM-I third year physics course vs the modelling-based third year physics course) an in-depth analysis was completed on NM-I scores of those students who in the third year took NM-I vs the modelling-based physics courses. In addition, the lowest quartile of scientific reasoners at the end of the first year of modelling was tracked to determine the effect of repeated exposure to MI on SRS over three years of exposure to MI.

To determine the relationship between scientific reasoning ability and content knowledge in physics, chemistry, and biology modelling-based courses two regression analyses were conducted for each year of science instruction. One regression had the post-SR score as the dependent variable with prior SR and content scores as the independent variables. The other regression had post-content test scores as the dependent variable with prior SR and content scores as the independent variables.

RESULTS

All Ninth-Graders Have Similar Academic Abilities at Start of the Study

Graphs of the mean performance on the ERB CTP and ISEE with SEM bars (**Figure 5** and **Figure 6**) show that the students in the modelling sequence and the non-modelling inquiry sequence were very similar academically; having mean performances on multiple academic measures that were within the standard error of measurement. When differences were suggested (e.g., writing mechanisms), MI sequence performed lower than NM-I sequence.

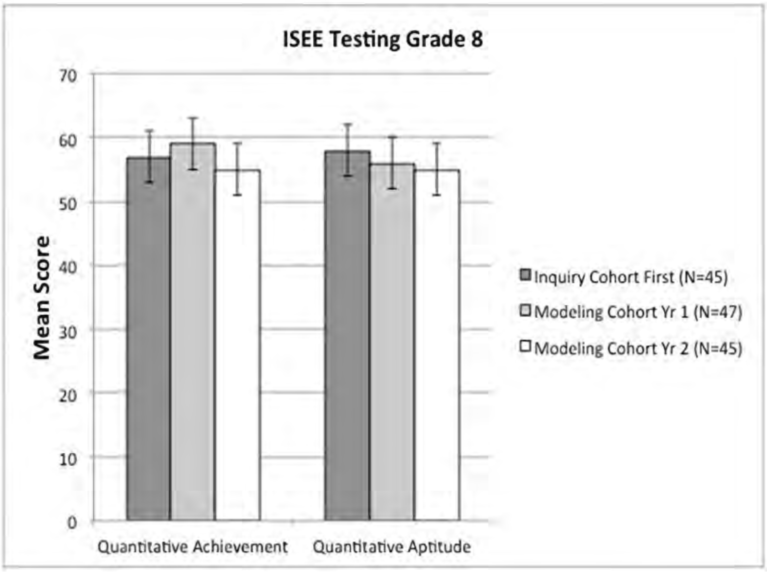

Figure 6. Mean ISEE grade 8 test scores for quantitative achievement & quantitative aptitude (Source: Authors)

Table 4. Fixed effect regression coefficients of HLM on scientific reasoning test scores

	Model 0	Model 1	Model 2	Model 3
Score level variable				
Intercept		57.8*** (0.9)	59.3*** (0.9)	59.0*** (1.7)
Time (in months)		0.2*** (0.03)	0.4*** (0.03)	0.2*** (0.04)
Modelling-based instruction		9.3*** (1.0)		9.9*** (1.0)
Student level variable				
First year modelling-based instruction				-2.74 (NS) (2.0)
Estimate variance components				
Deviance	9,853	9,541	9,642	9,539
Estimated parameters	2	2	2	2

Note. Standard deviations are shown in parentheses& ***p<.001

Longitudinal Analysis of Scientific Reasoning Scores

Longitudinal HLM analysis was conducted on the scientific reasoning scores of all students with scores nested within students (Raudenbusch & Byrk, 2002). Exposure to modelling-based pedagogy and timing of modelling-based pedagogy (first or third year) were included as independent variables, and time as a covariate. Students who only had one scientific reasoning score or students who did not have an initial scientific reasoning score were eliminated. All assumptions (e.g., homoscedasticity, normality, independence) were met. Because MI year 1 and MI year 2 did not differ significantly on their pretest means nor their longitudinal progression as shown in HLM analysis, these two groups were combined into one modelling-based group that received modelling pedagogy.

There was a significant effect (9.3%-9.9% over the course of the study, $p<0.001$) of MI on scientific reasoning test scores (**Table 4**, model 1 and model 3). There was also a significant effect of time on scientific reasoning test scores of approximately 0.2% per month when MI was included as a variable, which equates to 6.6% over the course of the study. It is worth noting that the timing of physics Instruction (i.e., first or third year of a three-year science sequence) did not have a significant impact on scientific reasoning. When a dummy variable was included that indicated whether students had taken MI in their first year (1) or their third year (0), this variable was not significant (first year MI, model 3 in **Table 4**).

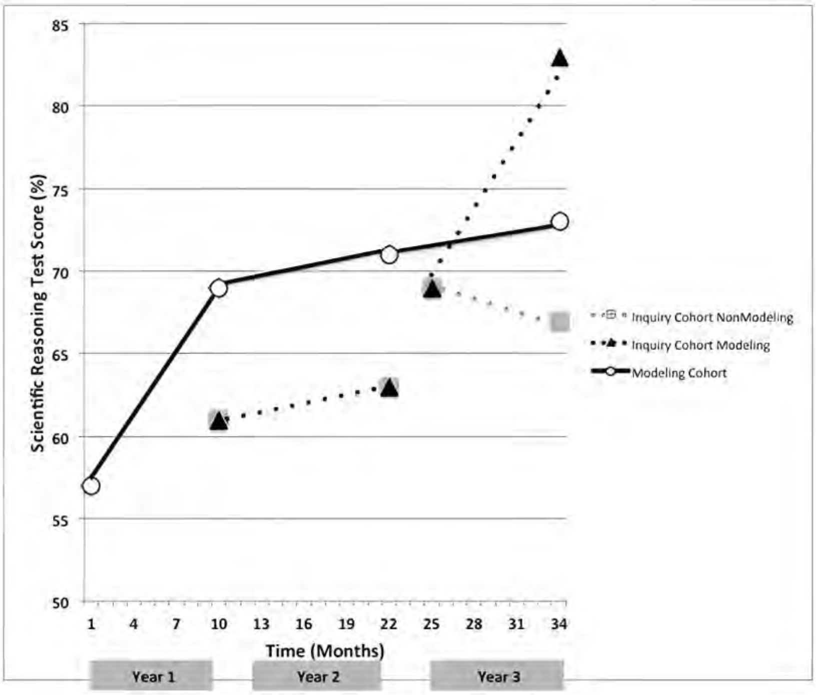

Figure 7. Mean scientific reasoning test scores over time for modeling & inquiry cohorts (Source: Authors)

Comparison of Scientific Reasoning Scores for Modelling and Non-Modelling Inquiry Sequences

Graphing the mean scientific reasoning scores for the modelling sequence over time (**Figure 7**) revealed that scores increased significantly from 57.0% to 69.0% during the first exposure to modelling in physics (repeated measures ANOVA, F (1, 248)=191, p<.001, Cohen's d=0 .9). 69.0% mean scientific reasoning score achieved by MI sequence at the end of their first year was significantly greater than that achieved by NM-I sequence at the end of its first year (M=62.0%, ANOVA, F (1, 337)=19, p<.001, Cohen's d=0.5).

At 33 months, there is a significant difference across all three conditions (F(2, 255)=6.2, p=.002). Post-hoc analysis with a Bonferroni correction, reveals that the only significant difference is between inquiry cohort modelling (83.0%) and inquiry cohort non-modelling (67.0%, p=.002, Cohen's d=.8). Both inquiry cohort modelling and inquiry cohort non-modelling are not statistically significantly different from modelling cohort (73.0%, n=190, p=.05 and p=.09, respectively), but the effect sizes (Cohen's d=.6, Cohen's d=0.4, respectively) suggests that the failure to detect significance may be due to the small sample sizes for inquiry cohort modelling (n=25) and inquiry cohort non-modelling (n=43). The difference in sample sizes between the two inquiry cohorts and modelling cohort is caused by differences in course selection. Inquiry cohort could opt into taking physics whereas all students needed to take biology. Possible reasons to explain SRS results at 33 months will be considered later.

Because time influences scientific reasoning scores, as time increases there is less room for improvement on LCTSR. To compensate for this decrease, instead of straight gains, Hake's (1998) normalized gains, which calculates the amount of gain as a proportion of the possible gain, were calculated for each student each year. As seen in **Figure 8**, the average normalized gain across students for modelling-based pedagogy in the initial year is high, approximately 26.0% (as was suggested by HLM analysis). The effect of modelling in subsequent years for the modelling sequence appears to level off producing nonsignificant changes in scores as reflected in the shallow slopes in **Figure 7** and SEM bars that overlap with zero in **Figure 8**, year 2 and year 3).

However, the initial difference in scientific reasoning scores achieved at the end of the first year of science instruction between NM-I and MI sequences was still maintained at the end of the second year (**Figure 7**, M_M=71, M_I=63).

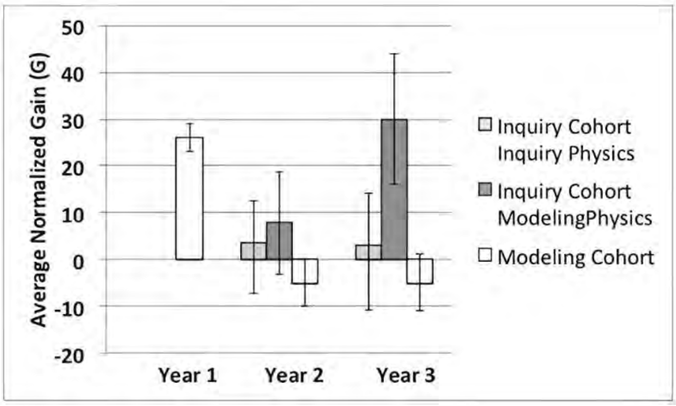

Figure 8. Hake's (1998) normalized gains of scientific reasoning scores averaged across students by year, cohort, & exposure to modeling instruction with SEM bars (Source: Authors)

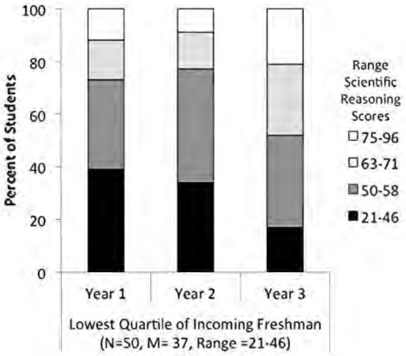

Figure 9. Change in distribution of scientific reasoning scores on end-of-year post-tests for the lowest quartile of incoming freshmen (Source: Authors)

Modelling-Based Instruction in Third Year of NM-I Cohort Produces Gains in Scientific Reasoning Scores

In NM-I sequence, students took a biology, chemistry, and physics sequence but not all opted to take physics in the third year. This is indicated in **Figure 7** by the absence of lines connecting the second and third years of science instruction. At the beginning of the third year of science instruction, there appears to be a large jump in scientific reasoning scores for NM-I sequence. Those who did opt to take physics as a third year of science tended to score slightly higher on the scientific reasoning test at the end of their second year of high school science (n=61, M=65) than those who did not (n=23, M=57).

Within the group that opted to take physics as a third year of science, students were randomly placed into one of two teachers' classrooms. One teacher taught physics using modelling-based pedagogy while the other used a mix of inquiry and scientific discovery methods. In the second year of instruction, the average Normalized gains in scientific reasoning of these two subsets of the cohort were not significantly different from zero (**Figure 8**, year 2). However, in the third year of science instruction, those students who received MI in physics achieved average normalized gains of approximately 30.0%, while those students who received inquiry/discovery instruction in physics showed average normalized gains that were not significantly different than zero (**Figure 8**, year 3). Therefore, modelling has a nonzero effect on normalized gains in scientific reasoning scores regardless of the year of instruction reinforcing the conclusion from HLM analysis.

Repeated Exposure to Modelling-Based Instruction Benefits Lower Performing Students

In MI sequence, there is an apparent leveling off of scientific reasoning scores after the first year. However, this large-scale analysis may be missing within-group effects of continued MI. Therefore, we examined the effect of continued MI on students within the cohort most likely to be at risk–those in the lowest quartile of scientific reasoning scores at the beginning of their first year of science instruction. **Figure 9** shows a shift in

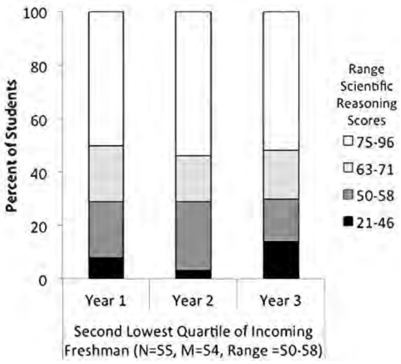

Figure 10. Change in distribution of scientific reasoning scores on end-of-year post-tests for the second lowest quartile of incoming freshmen (Source: Authors)

Table 5. Means & standard deviations (in parentheses) for content tests & scientific reasoning tests for MI cohort used in linear regression analyses

Year	Year 1		Year 2		Year 3	
Test	Pre	Post	Pre	Post	Pre	Post
SR	55 (18) & n=136	69 (16.8) & n=137	--	69 (16.7) & n=106	--	74 (16.4) & n=102
FCI	22 (8.9) & n=137	45 (15.3) & n=134	--	--	--	--
CCI	--	--	--	41 (16.7) & n=119	--	43 (18) & n=104
BioSatII	--	--	--	--	36 (11.3) & n=110	53 (12.4) & n=114

Table 6. Standardized beta coefficients for significant variables in regression models

IV	PreSR year 1	PostSR year 1	PostSR year 2	PreFCI year 1	PostFCI year 1	PostCCI year 2	PreSATII year 3	Regression statistic
PostSR year 1	.65***	NI	NI	NS	NI	NI	NI	F(2,128)=57 & p<.001
PostSR year 2	NI	.59***	NI	NI	NS	NI	NI	F(2,100)=40 & p<.001
PostSR year 3	NI	NI	.60***	NI	NI	NS	NS	F(3,72)=23 & p<.001
PostFCI year 1	.29***	NI	NI	.25**	NI	NI	NI	F(2,128)=17 & p<.001
PostCCI year 2	NI	.38***	NI	NI	.23*	NI	NI	F(1,110)=21 & p<.001
PostCCI year 3	NI	NI	.19*	NI	NI	.56***	NS	F(2,71)=25 & p<.001
PoSATII year 3	NI	NI	.24*	NI	NI	.23*	.27*	F(3,78)=13 & p<.001

Note. ***p<0.001; **p<0.01; *p<0.05; NS: Not significant; NI: Not included; IV: Independent variable; FCI: Force concept inventory; & CCI: Chemistry concept inventory

the distribution of scientific reasoning scores toward higher performance for this quartile with continued MI. For comparison, **Figure 10** shows the distribution of scientific reasoning scores for the next highest quartile. No shift in distribution was seen, a pattern that was also observed with the top two quartiles.

Scientific Reasoning Scores are a Significant Predictor of Post-Content Test Scores

At the end of each course for the second year of the conversion to MI, content assessments were given as well as the LCTSR. These students took physics in their first year, chemistry in their second year, and biology in their third year of science instruction. The mean scores are shown in **Table 5**.

To assess the interplay between content knowledge and scientific reasoning, two regression analyses were conducted for each year of science instruction. One set of regression models had the post-SR score as the dependent variable with prior SR and content scores as the independent variables (**Table 6**).

The other set of regression models had post-content test scores as the dependent variable with prior SR and content scores as the independent variables. In both sets of regression analyses, “prior” refers to the most proximate previous scores. For example, at the end of the second year of science instruction, the prior SR reasoning score was that given at the end of the first year of science instruction, and the prior content test score was FCI also given at the end of the first year of science instruction. For every year, post-SR score was shown to be affected by the prior SR score, but not the prior content score. Meanwhile, post-content score was shown to be predicted by both the prior content score and the prior SR score. These results suggested that students’ SR influenced their ability to do well on content assessments, but the reverse is not true.

DISCUSSION OF RESULTS

This longitudinal case study was undertaken to determine the effects on the scientific reasoning of students when exposed to Modelling-based pedagogy over three years of science study. The use of LCTSR was a strategic choice since it is a well-respected and validated instrument that would allow for additional comparisons to other contexts thus extending the generalizability of the results. Five main findings emerged from the analysis of the three-year longitudinal study.

First-Year Modelling-Based Instruction Students Had Higher Scientific Reasoning Skills and Maintained Them Across All Three Years of Study

First-year students were similar in all respects across multiple areas as shown by EERB and ISEE results (see **Figure 5** and **Figure 6**, respectively) and when there were differences, MI cohort performed lower than the non-modelling inquiry cohort. Thus, differences in reasoning skills at the end of the first year are unlikely to be attributable to initial differences between MI and NM-I cohorts. Despite this cognitive similarity between the two entering cohorts, students in MI sequence compared to NM-I sequence scored much higher on scientific reasoning at the end of the first year (69.0% vs. 57.0%, respectively).

Compared to NM-I cohort, MI cohort maintained these higher scores through the end of the second year of study (71.0% vs. 63.0%). Thus, continued improvement of average scientific reasoning scores due to a second year of modelling was not seen. During the third year, MI cohort continued to improve at a rate of 0.2% per month. Thus, MI cohort demonstrated a final score of 73.0% on average. These results mean that the students in MI cohort achieved scientific reasoning scores at the end of their first year of high school well above the prescientific reasoning scores reported for non-STEM undergraduate college majors (54.0%) and very close to those reported for undergraduate STEM majors (75.0%) (e.g., Moore & Rubbo, 2012). Thus, by the end of their third year, the average MI student demonstrated scientific reasoning abilities equivalent to those of incoming STEM majors even though most of these students would be moving into non-STEM related fields. The use of modelling-based pedagogy as a science practice across years, therefore, portends well for supporting college readiness. This is especially good news given that SRS have been shown to resist changes during college years (e.g., Ding, 2018; Ding et al., 2016). This longitudinal study further supports earlier short intervention studies that found gains in SRS when pedagogies that focus on critical thinking skills were deployed s (e.g., Blumer & Beck, 2019; Hester et al., 2018; Yanto et al., 2019).

A comparison of reasoning scores between MI and NM-I cohorts at the end of the third year was attempted and demonstrated no significant differences. However, it is our contention that it is not fruitful to compare the scientific reasoning scores between the cohorts in the third year for multiple reasons. One reason is that the third-year course selection choices for NM-I cohort included multiple avenues (see **Figure 3**). These alternate paths caused a decrease in the number of students in each path lowering the power of the analysis, which might be one reason that no reasoning scores in the third year were significantly different even though the averages were very different ranging from the mid 60's to the mid 80's (see **Figure 7**). In addition, the NM-I students who chose science in the third year were a select group of students planning on continuing to enroll in future science classes. The removal of the students from NM-I cohort who did not plan on enrolling in further science classes would have affected the overall average in the third year probably negatively. On the other hand, MI third-year cohort included all students in the school as it was a graduation requirement (i.e., those planning on pursuing future science classes as well as those who were not). Thus, any comparisons of these two cohorts scores at the end of the third year would not be valid.

Content Post-Test Scores Were Correlated to Pre-Scientific Reasoning Scores

Significant correlations between prescientific reasoning scores and content post-test scores in multiple subjects were observed in this study. But the reverse was not true. This is in line with prior research studies, which showed that gains in scientific reasoning scores were correlated with conceptual gains at the college and high school levels (e.g., Nieminen et al., 2012; Orosz et al., 2023; Thompson et al., 2018). Thus, MI cohort's science reasoning pre-scores correlated with larger gains in their subject knowledge as the years progressed.

Scientific Reasoning Increases Are Due to Modelling-Based Instruction and Not to the Switch in Discipline in the First Year of Study

The results indicated that the scientific reasoning scores were due to MI and not to the switch from biology to physics in the first year of secondary school. HLM study revealed these interesting results when analyzing the non-modelling inquiry sequence during their third year of study. In the third-year students taking physics took one of two paths: either a MI or a NM-I course. The students in the physics NM-I option (with no exposure to modelling) demonstrated no significant normalized gain in their reasoning scores across the third year. However, NM-I cohort students assigned to the modelling-based physics course demonstrated an average 30.0% normalized gain in reasoning scores by the end of the third year. It is unlikely that this gain was due to teacher effect since the teacher who taught the third-year NM-I cohort's modelling-based physics course also taught a first-year MI physics course in the MI sequence (see **Table 1**). The scores of this teacher's first-year students on LCTSR were in the middle of the range of first-year students. Combined, these results provided support that the differences between NM-I cohort and MI cohort's scientific reasoning results cannot be caused by simply switching from biology to physics in the freshman year. Thus, the data suggested that MI affects the students' scientific reasoning scores no matter at what time the course is taken and implies that physics-first is not the magic bullet to increasing scientific reasoning scores, MI is. This finding further supports the use of pedagogies that focus on critical thinking skills to enhance SRS at all levels (e.g., Blumer & Beck, 2019; Hester et al., 2018; Yanto et al., 2019).

Timing Effects of First Exposure to Modelling-Based Instruction Support MI Cohort's Content Scores

The zero normalized gains and low slopes for continued MI in MI cohort suggested that after the first exposure to modelling, little benefit was obtained with more exposure. This brings up the question: does it matter if students experience modelling throughout the course of their high school years or is a single exposure sufficient? Our regression analyses with content knowledge and scientific reasoning would argue against a capstone exposure in the last year of science instruction even if the graduating scientific reasoning scores are not significantly different from those obtained by first-year exposure. As shown here and in prior research prescientific reasoning scores are predictive of post-year content scores, but the reverse is not true (e.g., Moore & Rubbo,2012). This suggests that scientific reasoning ability at the beginning of the year could be playing a role in content learning. This argument was supported by research on college students, which shows that students' prescientific reasoning scores are correlated with content gains (e.g., Getahun, 2022; Thompson et al., 2018; Williams et al., 2021). Theoretically, if scientific reasoning is responsible for increasing content knowledge, the earlier students are exposed to modelling, the greater the benefit for subsequent knowledge acquisition. However, there is an even stronger argument to be made not only for exposing students to MI early but for continuing that mode of instruction in subsequent years.

Repeated Modelling Exposure Benefits Lower Performing Students

While overall there seems to be a leveling off of the scientific reasoning scores for MI students on average this is not the case when one looks at the scores of the students at most risk of performing poorly (the lowest quartile of students). The lowest quartile of incoming ninth-graders in MI sequence showed a marked decrease in the number of reasoning scores below 40.0% over the three years of the study. Thus, they experienced a concomitant increase in reasoning scores above 75.0% (see **Figure 9** and **Figure 10**). Given the correlation shown between content learning and SRS, this finding demonstrates that MI could allow the students at most risk in science classes to thrive and be better prepared for their future. Thus, science modelling could be important for promoting equity in school science. This finding is supported by a cross-national study's findings concerning the differential effectiveness of classroom-level factors like modelling on science and mathematics achievement. Vanalar et al. (2016) determined that these classroom factors seemed to impact the achievement levels of lower performing students to a greater extent. Similarly, in our study, we determined that one impact of long-term exposure to scientific modelling is that it produces higher SRS in the lowest-performing students. Thus, one year of science modelling might not be enough for these students to succeed at their utmost. This to our knowledge has not been reported elsewhere and is one that would not have been discovered without the longitudinal nature of this research.

The importance of this finding cannot be stressed enough especially in terms of future persistence in science. Past studies have shown that students' beliefs in their competency are predictive of achievement (Khan et al., 2022; Schunk et al., 2008).

CONCLUSIONS

The findings presented in this paper show that in terms of the conceptual framework used in this study, MI supports the development of SRS, which support student content learning. Thus as a coherent whole, the evidence presented in this paper argues for sustained MI in high school science curricula. This study demonstrated that there is a benefit to starting MI in high school early and continuing exposure throughout secondary science instruction. The following evidence supports this statement:

1. MI had higher overall scientific reasoning scores than NM-I.
2. Scientific reasoning appears to positively correlate with content scores but not vice versa.
3. Consistent MI moved weaker students to higher scientific reasoning scores'.

Thus, there appears to be a benefit to starting MI in high school early and continuing it throughout science instruction.

The long-term exposure to MI in high school can continue to benefit students for years to come. Research implies that lower levels of scientific reasoning correspond with difficulties in problem-solving in abstract contexts such as constructing conclusions based upon evidence. Moving students towards higher levels of scientific reasoning better equip them as everyday citizens to develop evidence-based conclusions, which is particularly important as the world becomes more and more data-driven. For students moving on to college, improved reasoning scores set them up for success. Studies have shown that entry-level college students planning on majoring in STEM fields routinely score in the seventies whereas non-STEM majors routinely score in the fifties (e.g., Jensen et al., 2015; Moore & Rubbo, 2012). Because of the link shown in college between entry-level scientific reasoning scores and content gains (e.g., Ding, 2014; Moore & Rubbo, 2012) increasing scientific reasoning scores in high school should better equip students for college success, and thus has the potential to ameliorate the cycle of failure seen in entry-level college science courses (Jensen et al., 2015) as well as help to decrease the dropout rates of STEM majors (Heilbronner, 2011).

Thus, the marked shift of the lowest quartile of modelling sequence to higher levels of scientific reasoning suggests that continued MI not only better prepares these at-risk students to master new scientific content, but also better prepares them to succeed as everyday citizens and college students. It is highly probable given the lack of reasoning gains by NM-I cohort across the three years of the study that if this group of students had only participated in the initial year of modelling their scientific reasoning scores would have remained below 50.0% by the end of their third year of science instruction. This scoring range would put them well below entering undergraduate STEM majors and at the lower end of entering undergraduate non-STEM majors. Thus, using modelling as a scientific practice across multiple science classes and years allows for increased equity across all ranges of student abilities.

Limitations

This study is a long-term comparative case study of a single school implementing a change in pedagogy. Thus, while the results have the advantage of being obtained in a real-world context, the study is subject to several limitations. Administrative policy changes made direct gain score comparisons for scientific reasoning impossible for the first year of the study. However, given the comparability of entering students' scores on several measures, it is a reasonable supposition that NM-I and MI cohorts' pre-freshmen scientific reasoning scores would have been similar. A second limitation was that the study took place at a single independent school, which limits the generalizability to other contexts. However, the single location minimized the variability of context (in terms of non-science instruction and science teachers) and initial differences between all students. The final limitation is that the study had two independent variables including the differences in course pedagogy as well as a switch in the type of science discipline being taught at the freshman level. The results in the third-year physics course imply that it is not the science discipline sequence that makes the difference between the two cohorts' scientific reasoning but the use of modelling- based instruction. However, it is essential to test this hypothesis.

Implications and Future Directions

This research adds to the growing body of literature that indicates the importance of using modelling as scientific practice in secondary schools (Dye et al., 2013; Osborne, 2014). This research demonstrated the longitudinal effects of modelling-based science instruction. Our results suggest that the use of modelling as a scientific practice is a social justice as well as an equity issue. If schools wish to improve all students' scientific reasoning abilities to make them successful and thoughtful consumers of scientific knowledge as well as prepare them for STEM careers, then our results indicated that MI should not only begin as early as the ninth grade but should continue throughout high school. This longitudinal comparative case study lays the groundwork for additional longitudinal research that focuses on the use of modelling as scientific practice in a quasi-experimental fashion across multiple contexts to determine the generalizability of these findings.

Author contributions: All authors were involved in concept, design, collection of data, interpretation, writing, and critically revising the article. All authors approved the final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Ethics declaration: Authors declared that the research reported in the article was deemed to be exempt from the need for ethics since it made use of classroom artifacts. However, authors required that any student data used in the research had to have parent permission. Therefore, parent consent forms were collected from all parents of any students whose data was included in this study.

Declaration of interest: Authors declare no competing interest.

Data availability: De-identified data is available upon request to the corresponding author.

REFERENCES

Aakre, I. G., Persson, J. R., Lein, H. L., & Eggen, P. O. (2021). First-year university students' perception of chemical bonding and bond energy. *Nordic Journal of STEM Education, 4*(2), 12-30. https://doi.org/10.5324/njsteme.v4i2.3300

Al-Balushi, S. M., Al-Musawi, A. S., Ambusaidi, A. K., & Al-Hajri, F. H. (2017). The effectiveness of interacting with scientific animations in chemistry using mobile devices on grade 12 students' spatial ability and scientific reasoning skills. *Journal of Science Education and Technology, 26*(1), 70-81. https://doi.org/10.1007/s10956-016-9652-2

Bao, L., Fang, K., Cai, T., Wang, J., Yang, L., Cui, L., Han, J., Ding, L., & Luo, Y. (2009). Learning of content knowledge and development of scientific reasoning ability: A cross culture comparison. *American Journal of Physics, 77*(12), 1118-1123. https://doi.org/10.1119/1.2976334

Bao, L., Xiao, Y., Koenig, K., & Han, J. (2018). Validity evaluation of the Lawson classroom test of scientific reasoning. *Physical Review Physics Education Research, 14*(2), 020106. https://doi.org/10.1103/PhysRevPhysEducRes.14.020106

Barlow, A. T., Frick, T. M., Barker, H. L., & Phelps, A. J. (2014). Modeling instruction: The impact of professional development on instructional practices. *Science Educator, 23*(1), 14-26.

Berland, L. K., Schwarz, C. V., Krist, C., Kenyon, L., Lo, A. S., & Reiser, B. J. (2016). Epistemologies in practice: Making scientific practices meaningful for students. *Journal of Research in Science Teaching, 53*(7), 1082-1112. https://doi.org/10.1002/tea.21257

Bernard, P., & Dudek-Różycki, K. (2019). Influence of training in inquiry-based methods on in-service science teachers' reasoning skills. *Chemistry Teacher International*, *1*(2), 20180023. https://doi.org/10.1515/cti-2018-0023

Blumer, L. S., & Beck, C. W. (2019). Laboratory courses with guided-inquiry modules improve scientific reasoning and experimental design skills for the least-prepared undergraduate students. *CBE–Life Sciences Education, 18*(1), ar2. https://doi.org/10.1187/cbe.18-08-0152

Bouzid, T., Kaddari, F., & Darhmaoui, H. (2022). Force and motion misconceptions' pliability, the case of Moroccan high school students. *The Journal of Educational Research, 115*(2), 122-132. https://doi.org/10.1080/00220671.2022.2064802

Buckley, B. C., Gobert, J. D., Kindfield, A. C. H., Horwitz, P., Tinker, R. F., Gerlits, B., Wilensky, U., Dede, C., & Willett, J. (2004). Model-based teaching and learning with Biologica: What do they learn? How do they learn? How do we know? *Journal of Science Education and Technology, 13*, 23-41. https://doi.org/10.1023/B:JOST.0000019636.06814.e3

Cameron, K., Malone, K. L., Sabree, Z., & Schuchardt, A. (2023). Lazy lizards in a drought: Science modeling and English learners. *Science Activities*, 1-12. https://doi.org/10.1080/00368121.2023.2200918

Carleschi, E., Chrysostomou, A., Cornell, A. S., & Naylor, W. (2022). Probing the effect on student conceptual understanding due to a forced mid-semester transition to online teaching. *European Journal of Physics, 43*(3), 035702. https://doi.org/10.1088/1361-6404/ac41d9

Chang, C. Y. (2010). Does problem-solving=prior knowledge + reasoning skills in earth science? An exploratory study. *Research in Science Education, 40*(2), 103-116. https://doi.org/10.1007/s11165-008-9102-0

Clement, J. J., & Steinberg, M. S. (2002). Step-wise evolution of mental models of electric circuits: A "learning-aloud" case study. *The Journal of the Learning Sciences, 11*(4), 389-452. https://doi.org/10.1207/S15327809JLS1104_1

Coletta, V. P., & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, pre-instruction scores and scientific reasoning ability. *American Journal of Physics, 73*(12), 1172-1182. https://doi.org/10.1119/1.2117109

Coletta, V. P., Phillips, J. A., & Steinert, J. J. (2007). Why you should measure your students' reasoning ability. *The Physics Teacher, 45*, 235-238. https://doi.org/10.1119/1.2715422

College Board. (2020). *The SAT subject tests student guide*. https://satsuite.collegeboard.org/media/pdf/sat-subject-tests-student-guide.pdf

Ding, L. (2014). Verification of causal influences of reasoning skills and epistemology on physics conceptual learning. *Physical Review Special Topics–Physics Education Research, 10*(2), 023101. https://doi.org/10.1103/PhysRevSTPER.10.023101

Ding, L. (2018). Progression trend of scientific reasoning from elementary school to university: A large-scale cross-grade survey among Chinese students. *International Journal of Science and Mathematics Education, 16*(8), 1479-1498. https://doi.org/10.1007/s10763-017-9844-0

Ding, L., Wei, X., & Liu, X. (2016). Variations in university students' scientific reasoning skills across majors, years, and types of institutions. *Research in Science Education, 46*(5), 613-632. https://doi.org/10.1007/s11165-015-9473-y

Dori, Y. J., & Kaberman, Z. (2012). Assessing high school chemistry students' modeling sub-skills in a computerized molecular modeling learning environment. *Instructional Science, 40*, 69-91. https://doi.org/10.1007/s11251-011-9172-7

Dukerich, L. (2015). Applying modeling instruction to high school chemistry to improve students' conceptual understanding. *Journal of Chemical Education, 92*(8), 1315-1319. https://doi.org/10.1021/ed500909w

Dye, J., Cheatham, T., Rowell, G. H., Barlow, A. T., & Carlton, R. (2013). The impact of modeling instruction within the inverted curriculum on student achievement in science. *Electronic Journal of Science Education, 17*, 1-19.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research, 82*(3), 300-329. https://doi.org/10.3102/0034654312457206

Georgiou, H., & Sharma, M. (2020). Engaging science academics with evidence based practices: Use of concept inventories in chemistry and physics across eight universities. *International Journal of Innovation in Science and Mathematics Education, 28*(4), 28-43. https://doi.org/10.30722/IJISME.28.04.003

Getahun, D. A. (2022). Scientific reasoning among teachers and teacher trainees: The case in Ethiopian schools and teacher training colleges. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-022-10347-6

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science, 71*, 742-752. https://doi.org/10.1086/425063

Gobert, J. D., O'Dwyer, L., Horwitz, P., Buckley, B. C., Tal Levy, S., & Wilensky, U. (2011) Examining the relationship between students' understanding of the nature of models and conceptual learning in biology, physics, and chemistry. *International Journal of Science Education, 33*(5), 653-684. https://doi.org/10.1080/09500691003720671

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology & Philosophy, 21*, 725-740. https://doi.org/10.1007/s10539-006-9054-6

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64-74. https://doi.org/10.1119/1.18809

Halloun, I. A. (2007). Mediated modeling in science education. *Science and Education, 16*(7), 653-697. https://doi.org/10.1007/s11191-006-9004-3

Harrison, A. G., & Treagust, D. F. (2000). Learning about atoms, molecules, and chemical bonds: A case study of multiple-model use in grade 11 chemistry. *Science Education, 84*(3), 352-381. https://doi.org/10.1002/(SICI)1098-237X(200005)84:3<352::AID-SCE3>3.0.CO;2-J

Heilbronner, N. N. (2011). Stepping onto the STEM pathway factors affecting talented students' declaration of STEM majors in college. *Journal for the Education of the Gifted*, *34(*6), 876-899. https://doi.org/10.1177/0162353211425100

Hestenes D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141-158. https://doi.org/10.1119/1.2343497

Hestenes, D. (2010). Modeling theory for math and science education. In R. Lesh, C. R. Haines, P. L. Galbraith, & A. Harford (Eds.), *Modeling students' mathematical modeling competencies* (pp. 13-41). Springer. https://doi.org/10.1007/978-1-4419-0561-1_3

Hester, S. D., Nadler, M., Katcher, J., Elfring, L. K., Dykstra, E., Rezende, L. F., & 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

Illes, M., Wilson, P., & Bruce, C. (2019). Forensic epistemology: Testing the reasoning skills of crime scene experts. *Canadian Society of Forensic Science Journal, 52*(4), 151-173. https://doi.org/10.1080/00085030.2019.1664260

Jenkins, J. L., & Howard, E. M. (2019). Implementation of modeling instruction in a high school chemistry unit on energy and states of matter. *Science Education International, 30*(2), 97-104. https://doi.org/10.33828/sei.v30.i2.3

Jensen, J. L., Neeley, S., Hatch, J. B., & Piorczynski, T. (2015). Learning scientific reasoning skills may be key to retention in science, technology, engineering, and mathematics*. Journal of College Student Retention: Research, Theory & Practice, 19*(2), 126-144. https://doi.org/10.1177/1521025115611616

Kaygisiz, G. M., Gurkan, B., & Akbas, U. (2018). Adaptation of scientific reasoning scale into Turkish and examination of its psychometric properties. *Educational Sciences: Theory and Practice, 18*(3), 737-757.

KMK. (Ed.). (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss* [*Biology education standards for the Mittlere Schulabschluss*]. Wolters Kluwer.

Kuhn, D., & Dean, Jr, D. (2004). Metacognition: A bridge between cognitive psychology and educational practice. *Theory into Practice, 43*(4), 268-273. https://doi.org/10.1207/s15430421tip4304_4

Lawson, A. E. (1992). What do tests of "formal" reasoning actually measure? *Journal of Research in Science Teaching, 29*(9), 965-983. https://doi.org/10.1002/tea.3660290906

Lawson, A. E. (2004). The nature and development of scientific reasoning. *International Journal of Science and Mathematics Education, 2*(3), 307-338. https://doi.org/10.1007/s10763-004-3224-2

Lawson, A. E., Banks, D. L., & Logvin, M. (2007). Self-efficacy, reasoning ability and achievement in college biology. *Journal of Research in Science Teaching, 44*(5), 706-724. https://doi.org/10.1002/tea.20172

Lawson, A. E., Clark, B., Meldrum, E. C., Falconer, K. A., Sequist, J. M., & Kwon, Y. J. (2000). Development of scientific reasoning in college biology: Do two levels of general hypothesis-testing skills exist? *Journal of Research in Science Teaching, 37*(1), 81-101. https://doi.org/10.1002/(SICI)1098-2736(200001)37:1<81::AID-TEA6>3.0.CO;2-I

Lehrer, R., & Schauble, L. (2006). Scientific thinking and science literacy. In K. A. Renninger, I. E. Sigel, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology: Child psychology in practice* (pp. 153-196). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470147658.chpsy0405

Lehrer, R., & Schauble, L. (2012). Seeding evolutionary thinking by engaging children in modeling its foundations. *Science Education, 96*(4), 701-724. https://doi.org/10.1002/sce.20475

Liang, L. L., Fulmer, G. W., Majerich, D. M., Clevenstine, R., & Howanski, R. (2012). The effects of a model-based physics curriculum program with a physics first approach: A causal-comparative study. *Journal of Science Education and Technology, 21*(1), 114-124. https://doi.org/10.1007/s10956-011-9287-2

Malone, K. (2008). Correlations among knowledge structures, force concept inventory, and problem-solving behaviors. *Physics Review Special Topics Physics Education Research*, *4*, 020107. https://doi.org/10.1103/PhysRevSTPER.4.020107

Malone, K. L., & Schuchardt, A. (2016, January). The efficacy of modeling instruction in chemistry: A case study. In *Proceedings from HICE 2016: The 14th Annual Hawaii International Conference on Education* (pp. 1513-1518). Honolulu, HI.

Malone, K. L., & Schuchardt, A. (2020). Population growth modelling simulations: Do they affect the scientific reasoning abilities of students? In H. C. Lane, Sl. Zvacek, & J. Uhomoibhi (Eds.), *Computer supported education: 11th International Conference, CSEDU 2019,* Heraklion, Crete, May 2-4, 2019, *Revised Selected Papers*, in Communications in Computer and Information Sciences Series (Vol. 1022, pp. 285-307). Springer Nature. https://doi.org/10.1007/978-3-030-58459-7_14

Malone, K. L., Schuchardt A. M., & Sabree, Z. (2019). Models and modeling in evolution. In U. Harms & M. J. Reiss (Eds.), *Evolution education re-considered: Understanding what works* (pp. 207-226). Springer International Publishing. https://doi.org/10.1007/978-3-030-14698-6

Malone, K. L. (2023). The effects of modeling-based pedagogy on conceptual understanding, scientific reasoning skills, and attitudes towards science of English Learners. *Science Education*, 1-33. https://doi.org/10.1002/sce.21805

Mehl, C. E. (2022). *Student experience and outcomes of chemistry modeling instruction* [Doctoral dissertation, The Ohio State University].

Ministry of Education in Taiwan. (2018). Curriculum standards for grades 1-12. *Ministry of Education*. https://www.naer.edu.tw/

Ministry of Education. (2014). *Folkeskoleloven* [*The Education Act*]. https://www.retsinformation.dk/forms/r0710.aspx?id=176327

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction–What is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 474-496. https://doi.org/10.1002/tea.20347

Moore, J. C., & Rubbo, L. J. (2012). Scientific reasoning abilities of nonscience majors in physics-based courses. *Physical Review Special Topics–Physics Education Research, 8*(1), 10106. https://doi.org/10.1103/PhysRevSTPER.8.010106

Mulford, D. R., & Robinson, W. R. (2002). An inventory for alternate conceptions among first-semester general chemistry students. *Journal of Chemical Education, 79*(6), 739-744. https://doi.org/10.1021/ed079p739

National Research Council. (2015). Guide to implementing the next generation science standards. *Board on Science Education, Division of Behavioral and Social Sciences and Education*. https://nap.nationalacademies.org/catalog/18802/guide-to-implementing-the-next-generation-science-standards

NGSS Lead States. 2013. *Next generation science standards: For states, by state*s. The National Academies Press.

Nieminen, P., Savinainen, A., & Viiri, J. (2012). Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning. *Physical Review Special Topics–Physics Education Research, 8*(1), 010123. https://doi.org/10.1103/PhysRevSTPER.8.010123

Norris, S. P., Phillips, L. M., & Korpan, C. A. (2003). University students' interpretation of media reports of science and its relationship to background knowledge, interest, and reading difficulty. *Public Understanding of Science, 12*(2), 123-145. https://doi.org/10.1177/09636625030122001

OECD. (2018). *PISA 2018 results (volume I): What students know and can do*. PISA, OECD Publishing. https://doi.org/10.1787/5f07c754-en

Opitz, A., Heene, M., & Fischer, F. (2017) Measuring scientific reasoning–a review of test instruments. *Educational Research and Evaluation, 23*(3-4), 78-101. https://doi.org/10.1080/13803611.2017.1338586

Orosz, G., Németh, V., Kovács, L., Somogyi, Z., & Korom, E. (2023). Guided inquiry-based learning in secondary-school chemistry classes: A case study. *Chemistry Education Research and Practice, 24*(1), 50-70. https://doi.org/10.1039/D2RP00110A

Osborne, J. (2014). Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education, 25*(2), 177-196. https://doi.org/10.1007/s10972-014-9384-1

Passmore, C., & Stewart, J. (2002). A modeling approach to teaching evolutionary biology in high schools. *Journal of Research in Science Teaching, 39*(3), 185-204. https://doi.org/10.1002/tea.10020

Passmore, C., Stewart, J., & Cartier, J. (2009). Model-based inquiry and school science: Creating connections. *School Science and Mathematics, 109*(7), 394-402. https://doi.org/10.1111/j.1949-8594.2009.tb17870.x

Penuel, W. R., Fishman, B. J., Cheng, B. H., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design*. Educational Researcher, 40*(7), 331-337. https://doi.org/10.3102/0013189X11421826

Posthuma-Adams, E. (2014). How the chemistry modeling curriculum engages students in seven science practices outlined by the college board. *Journal of Chemical Education, 91*(9), 1284-1290. https://doi.org/10.1021/ed400911a

Raudenbusch, S. W., & Byrk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. SAGE.

Russ, R. S., & Odden, T. O. B. (2017). Intertwining evidence-and model-based reasoning in physics sensemaking: An example from electrostatics. *Physical Review Physics Education Research, 13*(2), 020105. https://doi.org/10.1103/PhysRevPhysEducRes.13.020105

Russ, R. S., Coffey, J. E., Hammer, D., & Hutchison, P. (2009). Making classroom assessment more accountable to scientific reasoning: A case for attending to mechanistic thinking. *Science Education, 93*(5), 875-891. https://doi.org/10.1002/sce.20320

Sapia, P., Napoli, F., & Bozzo, G. (2022). The Lawson's test for scientific reasoning as a predictor for university formative success: A prospective study. *Education Sciences, 12*(11), 814. https://doi.org/10.3390/educsci 12110814

Schinka, J. A., Velicer, W. F., & Weiner, I. B. (2003). *Handbook of psychology: Research methods in psychology*. John Wiley & Sons, Inc.

Schunk, D. H., Meece, J. R., & Pintrich, P. R. (2012). *Motivation in education: Theory, research, and applications*. Pearson.

Schwarz, C. V., 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

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

Thompson, E. D., Bowling, B. V., & Markle, R. E. (2018). Predicting student success in a major's introductory biology course via logistic regression analysis of scientific reasoning ability and mathematics scores. *Research in Science Education, 48*(1), 151-163. https://doi.org/10.1007/s11165-016-9563-5

Vanlaar, G., Kyriakides, L., Panayiotou, A., Vandecandelaere, M., McMahon, L., De Fraine, B., & Van Damme, J. (2016). Do the teacher and school factors of the dynamic model affect high-and low-achieving student groups to the same extent? A cross-country study. *Research Papers in Education, 31*(2), 183-211. https://doi.org/10.1080/02671522.2015.1027724

Williams, K. R., Wasson, S. R., Barrett, A., Greenall, R. F., Jones, S. R., & Bailey, E. G. (2021). Teaching Hardy-Weinberg equilibrium using population-level Punnett squares: Facilitating calculation for students with math anxiety. *CBE–Life Sciences Education, 20*(2), ar22. https://doi.org/10.1187/cbe.20-09-0219

Yanto, B. E., Subali, B., & Suyanto, S. (2019). Improving students' scientific reasoning skills through the three levels of inquiry. *International Journal of Instruction, 12*(4), 689-704. https://doi.org/10.29333/iji.2019.12444a

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*(1), 99-149. https://doi.org/10.1006/drev.1999.0497

◆ ❖ ◆