Mapping undergraduate chemistry students' epistemic ideas about models and modeling

Katherine Lazenby | Avery Stricker | Alexandra Brandriet | Charlie A. Rupp | Kathryn Mauger-Sonnek | Nicole M. Becker

University of Iowa, Iowa City, Iowa

Correspondence
Nicole M. Becker, University of Iowa, Iowa City, IA.
Email: nmbecker@uiowa.edu

Funding information
Division of Undergraduate Education, Grant/Award Number: 1611622

Abstract
Developing and using scientific models is an important scientific practice for science students. Undergraduate chemistry curricula are often centered on established disciplinary models, and assessments typically provide students with opportunities to use these models to predict and explain chemical phenomena. However, traditional curricula generally provide few opportunities for students to consider the epistemic nature of models and the process of modeling. To gain a sense of how introductory chemistry students understand model changeability, model multiplicity, the evaluation of models, and the process of modeling, we use a construct-mapping approach to characterize the sophistication of students' epistemic knowledge of models and modeling. We present a set of four related construct maps that we developed based on the work of other scholars and empirically validated in an undergraduate introductory chemistry setting. We use the construct maps to identify themes in students' responses to an open-ended survey instrument, the models in chemistry survey, and discuss the implications for teaching.

KEYWORDS
chemical education research, epistemic knowledge, metamodeling, models and modeling, undergraduate
INTRODUCTION

Scientific models and the practice of modeling are central to contemporary research in chemistry and other science disciplines because models are important and useful tools for explaining and predicting natural phenomena (Schwarz, Passmore, & Reiser, 2016). Commonly, general chemistry students are asked to use models such as molecular-level formulas, Lewis structures, and graphical and mathematical models as tools for problem solving (Harrison & Treagust, 2000). Moreover, developing and using models is one of the eight scientific practices highlighted by the National Research Council’s Framework for K-12 Science Education (2012) as a necessary component of authentic science curricula. The Framework argues that engaging students in scientific practices allows them to develop deeper knowledge of science content and cultivate an appreciation of the types of activities that constitute scientific inquiry.

Schwarz et al. (2016) conceptualized two ways that students might engage in the practice of developing and using models: thinking with models and thinking about models. Thinking with models, which has similarly been referred to as “model-based reasoning” (Windschitl, Thompson, & Braaten, 2008) and “learning and applying specific models” (Perkins & Grotzer, 2005), entails using models as tools for explaining or predicting phenomena. Consistent with the explanatory and generative nature of authentic scientific activity, students who engage in thinking with models generate explanations for how or why a natural phenomenon occurs and develop predictions about future outcomes. Typically, when students think with models, they use or apply previously established models. In traditional undergraduate chemistry curricula, for example, students may be asked to use an equilibrium constant expression to predict the extent to which a chemical reaction will proceed.

Though chemistry students may often be asked to think with models, from our perspective, traditional chemistry curricula less frequently require students to engage in thinking about models. Thinking about models is related to what has been called epistemic knowledge of models and modeling (Krell, Upmeier zu Belzen, & Krüger, 2014; Passmore, Gouvea, & Giere, 2014; Pluta, Chinn, & Duncan, 2011), which has also been referred to as metamodeling knowledge (Schwarz et al., 2016; Schwarz & White, 2005). By epistemic knowledge, we mean knowledge about what is knowable, how one comes to know things, and what counts as knowing (Duncan & Rivet, 2013); thus by extension, epistemic knowledge of models and modeling relates to knowledge about how models come to be, the purpose of models, and what counts as a model.

Schwarz et al. (2016) argued that engaging students in thinking about models, together with thinking with models, can “help students gain more ownership of ideas,” promote connections between theoretical ideas and real-world phenomena, and promote deeper content learning (p. 118). Thinking about models may also affect students’ understanding of science beyond formal education. From a scientific literacy perspective, the ability to think about models may support students’ ability to meaningfully evaluate socially relevant models (Berland et al., 2016; Greene & Yu, 2015; Justi & Gilbert, 2002; Schwarz et al., 2016, 2009).

Scholars and practitioners have developed some modeling-focused curricula for high school and undergraduate-level courses that aim to engage students in the practice of developing and using scientific models and thinking both with and about scientific models (e.g., Brewe, 2008; Hestenes, 1987; Schuchardt & Schunn, 2016; Schwarz et al., 2009; Tien, Rickey, & Stacy, 1999; Wells, Hestenes, & Swackhamer, 1995; Windschitl et al., 2008). Researchers have found that such model- and modeling-focused curricula increase science content learning compared to traditional lecture-based curricula (Brewe et al., 2010; Hestenes, 1987; Jackson, Dukerich, &
Hestenes, 2008; Wells et al., 1995). In addition, modeling-focused curricula may have positive effects on students’ attitudes toward science and problem-solving abilities (Brewe, Kramer, & O’Brien, 2009; McPadden & Brewe, 2017) and may narrow performance gaps between underrepresented minority students and white students (Brewe et al., 2010).

However, in both undergraduate chemistry and physics contexts there is limited evidence of how traditional or modeling-focused curricula impact students’ epistemic ideas about models and modeling (Justi & Gilbert, 2002; Nicolau & Constantinou, 2014). Prior efforts to assess metamodeling knowledge at the high school and undergraduate levels have focused used selected-response assessments for assessing the efficacy of curricular interventions on students’ perceptions of models and modeling (Burgin, Oramous, Kaminski, Stocker, & Moradi, 2018; Cheng et al., 2014; Cheng & Lin, 2015; Everett, Otto, & Luera, 2009; Gobert et al., 2011; Levy & Wilensky, 2009; Liu, 2006; Park, Liu, Smith, & Waight, 2017). For example, Treagust, Chittleborough, and Mamiala (2002) developed the Students’ Understanding of Models in Science (SUMS) instrument to assess students’ perceived metamodeling knowledge. SUMS, a 27-item Likert-response format survey to assess five dimensions of metamodeling knowledge: models as multiple representations, models as exact replicas, models as explanatory tools, uses of scientific models, and the changing nature of models. While this self-assessment has been used to assess the efficacy of modeling-focused interventions, some claims about students’ metamodeling knowledge based on such assessments are inconsistent with qualitative analyses of similar populations. For example, using the SUMS, Treagust et al. (2002) suggest that high school students have relatively robust understandings of metamodeling knowledge, while qualitative studies of similar populations report that students possess naïve and unsophisticated epistemic knowledge of models and modeling (Crawford & Cullin, 2005; Grosslight, Unger, Jay, & Smith, 1991; Grüenkorn, zu Belzen, & Krüger, 2014; Krell, Reinisch, & Krüger, 2015; Sins, Savelsbergh, van Joolingen, & van Hout-Wolters, 2009). An additional challenge is that because SUMS is a self-report, Likert-response format assessment, it has limited utility in assessing growth in metamodeling knowledge.

Since epistemic knowledge contributes to the development of content knowledge (Schuchardt & Schunn, 2016; Schwarz & White, 2005), we argue that the science community should develop ways to assess growth in students’ epistemic knowledge about models and modeling (Justi & Gilbert, 2002). While the literature describes some assessment resources (e.g., Treagust et al., 2002; Treagust, Chittleborough, & Mamiala, 2004), most of these have been developed and validated with reference to K-12 students’ and pre-service teachers’ understandings (Cheng & Lin, 2015; Derman & Kayacan, 2017; Everett et al., 2009; Gobert et al., 2011; Liu, 2006; Park et al., 2017). Thus, to better develop and refine modeling-focused instructional resources that support student learning, the science community needs assessment resources that are valid for undergraduate populations as well (Justi & Gilbert, 2002).

2 | RESEARCH QUESTION

Here, we describe our use of a construct-modeling approach to assessing undergraduate chemistry students’ understanding of four dimensions of epistemic knowledge about models and modeling. Our work was guided by the research question:

In what ways do students think about model changeability, model multiplicity, the evaluation of models, and the process of modeling?
In the present study, we used the research literature to inform the initial construction of hypothetical construct maps for four dimensions of metamodeling knowledge and then used survey data to refine and empirically validate these construct maps. We used the resulting construct maps to assess undergraduate chemistry students’ conceptions of models and modeling after one semester of university-level general chemistry.

The literature has shown that construct maps are productive tools for assessing the development of students’ knowledge and skills in chemistry (Becker, Noyes, & Cooper, 2016; Becker, Rupp, & Brandriet, 2017; Brandriet, Rupp, Lazenby, & Becker, 2018; Claesgens, Scalise, Wilson, & Stacy, 2009; Loertscher, Lewis, Mercer, & Minderhout, 2018; Sevian & Talanquer, 2014). We thus expect that the set of proposed construct maps will be useful for informing assessment of epistemic knowledge of modeling for both traditional and modeling-focused curricula. In the next sections, we discuss the prior research on students’ understanding of models and modeling at the undergraduate level and literature that informs our approach to construct map development.

3 LITERATURE BACKGROUND

3.1 Theoretical perspective: Resources framework

Our view of epistemic knowledge is informed by the resources framework (Hammer & Elby, 2003). From this perspective, students’ epistemic resources can be thought of as analogous to knowledge resources in conceptual change work that characterizes students’ conceptual structures as “knowledge in pieces” (diSessa, 1988, 2018). Hammer and Elby (2003) define epistemological resources as ideas about the development of knowledge that is constructed through students’ everyday experiences and which may be usefully applied in certain contexts. Through instruction and interaction with the world around them, students’ knowledge structure are reorganized and new epistemological resources may be added.

Importantly, the resources perspective argues that activation of epistemological resources may be context-dependent, which is consistent with previous studies on context-specificities of students’ metamodeling knowledge (e.g., Gobert et al., 2011; Krell et al., 2014). By recognizing students’ epistemological resources, Hammer and Elby argue that instructors can support scaffolding growth of new resources and facilitation of appropriate use of existing resources.

3.2 Prior research on dimensions of epistemic knowledge of models and modeling

In the current project, the construct of interest is students’ metamodeling knowledge, that is, students’ epistemic knowledge of models and modeling. The literature contains descriptions of several dimensions of metamodeling knowledge, including the nature of models, the purpose of models, model changeability, model multiplicity, the evaluation of models, and the process of modeling (Crawford & Cullin, 2005; Gogolin & Krüger, 2016, 2018; Grosslight et al., 1991; Grünkorn et al., 2014; Justi & Gilbert, 2002; Krell et al., 2015; Krell & Krüger, 2016, 2017; Schwarz & White, 2005; Van Der Valk, Van Driel, & De Vos, 2007). Researchers have described metamodeling knowledge as multidimensional and have examined several specific dimensions including model multiplicity, model changeability, model evaluation, and the process of modeling.
The construct of model multiplicity refers to the idea that there may be multiple models for a given phenomenon, either to highlight different features or to serve different purposes. In chemistry, multiple models are commonly used to represent and reason about different aspects of chemical phenomena and these different models may serve different functions, for instance, enabling mechanistic predictions about what will happen in the case of structural representations of molecules, versus making predictions using mathematical models. There is a broad literature base pertaining to specific ways in which use of multiple models or representations contributes to student understanding, and we direct the reader to Gilbert and Treagust (2009) for this discussion.

The construct of model changeability generally refers to the idea that models are human-constructed tools that can be and often are altered as a routine part of scientific activity. Evaluation of models refers to the idea that scientists can assess models' ability to explain and predict phenomena. Finally, we use the term “process of modeling” to refer to the idea that models are often based on patterns observed in empirical data or some combination of theory and data (Crawford & Cullin, 2005; Grosslight et al., 1991; Grünkorn et al., 2014; Schwarz & White, 2005).

Several scholars have discussed the qualitatively different ways that students (middle school through university [Crawford & Cullin, 2005; Grosslight et al., 1991; Grünkorn et al., 2014; Krell & Krüger, 2017; Schwarz & White, 2005]) and experts (Van Der Valk et al., 2007) think about models and modeling along these dimensions of metamodeling knowledge. For instance, Grosslight et al.’s foundational work in this area (1991) described students’ general epistemic knowledge of science based on interviews with two groups: 7th and 11th grade students (n = 33 and 22, respectively) who had not received any modeling instruction, and a group of adult experts (n = 4). Grosslight et al. proposed three levels of thinking about models based on interview findings: At Level 1, students understand models as toys or copies or scaled versions of reality that are useful for representing the target phenomenon as accurately as possible; At Level 2 students recognize that models are not exact replicas of reality, and are instead constructed for a purpose, with the modeler making choices about how to achieve the purpose; The focus remains, however, on representing reality rather than ideas. Level 3 represents the most expert-like ideas and includes the idea that the modeler takes an active role in constructing the model, and that models are constructed in the service of developing and testing ideas, rather than copying reality. Level 3 also included the idea of cycles of evaluation and model refinement based on data.

Grosslight et al. observed that most 7th grade students (67%) possessed Level 1 ideas across the constructs examined (model changeability, model multiplicity, designing and creating models, purpose of models, kinds of models) while 11th grade students possessed a mix of Level 1 and Level 2 ideas (23% pure Level 1, 36% pure Level 2, and 36% mixed Levels 1 and 2). None of the students expressed expert-like ideas (Level 3).

Building on the work of Grosslight and colleagues, Schwarz and White (2005) conducted a study with middle school physics students (N = 12 interviewees) who had received instruction on models and modeling in physics contexts (the Model Enhanced ThinkerTools curriculum). The authors used a three-level classification system, based in part on Grosslight et al.’s results (1991), to characterize students’ ideas about the definition of model, model multiplicity, model changeability, model evaluation criteria, the constructed nature of models, and the purpose of models. The authors observed that most students who had participated in the model-focused curriculum could provide moderate- to high-level explanations of the purposes of scientific models, the reasons for changing a model, and the
existence of multiple models, but offered low-level explanations of the evaluation of models as well as model creation and revision.

In another study building on the work of Grosslight et al. (1991), Crawford and Cullin (2005) characterized pre-service science teachers’ metamodeling knowledge before and after engaging with models- and modeling-focused learning modules. The authors developed a three- to four-level classification scheme for five constructs—the purpose of models, designing and creating models, model changeability, model multiplicity, and validating/testing models—based on relevant literature (Grosslight et al., 1991; Justi & Gilbert, 2002). Crawford and Cullin observed that before engaging with the learning modules, pre-service teachers exhibited mostly naïve metamodeling ideas across constructs; after engaging with the modules, most participants showed some increase in the sophistication of their metamodeling ideas but retained primarily low-level understandings of models and modeling.

Upmeier zu Belzen and Krüger (2010) developed three-level classification schemes for the constructs of the nature of models, the purpose of models, model changeability, model multiplicity, and testing models based on findings from previous studies (Crawford & Cullin, 2005; Grosslight et al., 1991; Justi & Gilbert, 2002). Although the manuscript is available only in German, other studies have translated and used their classification scheme, including Grünkorn et al. (2014) and Krell and Krüger (2017). In the former, the authors administered an open-ended survey to 1,177 German 7th to 10th graders to validate Upmeier zu Belzen and Krüger’s (2010) classification scheme with a large sample and observed that the three-level system worked well for the constructs of the nature and purpose of models but that an additional level, which they called “initial understanding,” was necessary to characterize students with the most naïve ideas about model multiplicity, model changeability, and testing models. For the constructs of testing models and model changeability, most students responded at a moderate level (Level 2), while for the constructs of the nature of models, the purpose of models, and model multiplicity, most students demonstrated a low level of understanding (Level 1).

Krell and Krüger (2017) used Upmeier zu Belzen and Krüger’s (2010) to assess university students’ understanding of three constructs: the purpose of models, testing models, and model changeability. The sample was composed of university students studying in STEM (science, technology, engineering, and math) and non-STEM (social sciences and linguistics) fields (N = 184) and an open-ended survey was the source of data. The authors found that few students possessed expert-like metamodeling knowledge according to Upmeier zu Belzen and Krüger’s (2010) classification scheme, and that students from STEM disciplines expressed more advanced metamodeling knowledge than students from non-STEM disciplines.

To identify experts’ perceptions of defining characteristics of models, Van Der Valk et al. (2007) administered a survey to practicing scientists who had recently published work related to scientific models or modeling in physics, astronomy, biology, chemistry, pharmacology, meteorology, geology, and interdisciplinary sciences journals (N = 24). Survey items included statements about the nature and purpose of models in scientific inquiry (e.g., “In your research work, two or even more different models can represent—and often do represent—the same object”). Participants indicated whether they believed the statement was consistent with the role that models played in their own research and provided an explanation for their choice. Overall, Van Der Valk and colleagues found that participants agreed with the provided statements and that participants’ descriptions of the role of models in their research provided useful insight about the role of models in contemporary research across disciplines.
3.3 | Comparison of qualitative levels of metamodeling knowledge

In Tables 1–4, we compare the qualitative levels described in each of the studies reviewed earlier, from the most naïve ideas to the expert or expert-like descriptions, for each construct (Crawford & Cullin, 2005; Grosslight et al., 1991; Grünkorn et al., 2014; Krell & Krüger, 2017; Schwarz et al., 2009; Van Der Valk et al., 2007).

3.4 | Construct modeling: A developmental perspective on student learning

In our present work on university students’ metamodeling knowledge, we have adopted Wilson’s (2004) construct-modeling approach. This perspective aligns with a constructivist and developmental perspective on student learning in which students’ knowledge develops along a continuum as their ideas change (Eylon & Linn, 1988; National Research Council, 2014).

In a construct-modeling approach, the target ideas or concepts of interest, in this case epistemological knowledge about models and modeling, are latent constructs that cannot be directly observed or measured. Instead, learners’ levels of knowledge about a given construct can be inferred from their responses to well-constructed assessments. Proponents of this approach assume that learners will progress through a series of qualitatively different levels ranging from the most naïve to the most advanced. Wilson (2004) describes four “building blocks” in developing assessments based on this construct-modeling approach. In this manuscript, we focus on the first “building block,” the development of construct maps.

A construct map is defined by Wilson (2009b, p. 3) as:

“a well thought out and researched ordering of qualitatively different levels of performance focusing on one characteristic. Thus, a construct map defines what is to be measured or assessed in terms general enough to be interpretable within a curriculum and potentially across curricula, but specific enough to guide the development of the other components.”

Researchers have developed and used construct maps as assessment tools in a variety of studies (e.g., Arya & Maul, 2012; Becker et al., 2016, 2017; Brandriet et al., 2018; Briggs, Alonzo, Schwab, & Wilson, 2006; N. J. Brown, Furtak, Timms, Nagashima, & Wilson, 2010; Claesgens et al., 2009; Loertscher et al., 2018; Rivet & Kastens, 2012; Schwarz et al., 2009; Sevian & Talanquer, 2014). Construct maps are distinct from scoring guides (sometimes referred to as rubrics), as scoring guides are typically developed for a single item, whereas construct maps provide a broader description of how learners’ thinking might theoretically progress (Wilson, 2004).

Though constructs are typically assumed to be unidimensional, this is not always the case; for multidimensional target constructs, each dimension can be treated as a separate construct (Briggs et al., 2006; Wilson, 2004). For example, Morell, Collier, Black, and Wilson (2017) used construct maps to assess four related dimensions (constructs) of students’ understanding of the nature of matter.

The development of construct maps often begins with the identification of upper and lower anchors. The upper anchor is based on existing standards or benchmarks, or, when such standards do not exist, the upper anchor can be defined by the normative or expert-like understanding. Lower anchors are based on the ideas that learners develop intuitively, without any
<table>
<thead>
<tr>
<th>Source</th>
<th>Most naïve ideas observed</th>
<th>Low-level</th>
<th>Moderate</th>
<th>Expert or expert-like</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grosslight et al. (1991)</td>
<td>Models cannot be changed</td>
<td>Models may change if something is wrong with model or how model was made</td>
<td>Models may change because new information is found</td>
<td>Changing a model is not only possible, it is inevitable</td>
</tr>
<tr>
<td>Schwarz and White (2005)</td>
<td>Model revision may not occur or would occur because the model was wrong</td>
<td>Model revision occurs when there is new information or evidence</td>
<td>Model revision may occur by rethinking one’s data and their implication as well as the purpose of the model</td>
<td></td>
</tr>
<tr>
<td>Grünkorn et al. (2014)</td>
<td>No reason for alterations</td>
<td>Model revision may occur in order to correct errors in the model object</td>
<td>Model revision may occur due to new findings about the original</td>
<td>Model revision may occur due to falsification of hypotheses about the original with the model</td>
</tr>
<tr>
<td>Krell and Krüger (2017)</td>
<td>Models are not changed.</td>
<td>A model is changed primarily when new discoveries are made.</td>
<td>A model is changed when it does not behave like the modeler wants it to.</td>
<td>Models are temporary in nature. A model is changed when its behavior is not in agreement with observations of the target.</td>
</tr>
<tr>
<td>Crawford and Cullin (2005)</td>
<td>Models are not changed.</td>
<td>A model is changed primarily when new discoveries are made.</td>
<td>A model is changed when it does not behave like the modeler wants it to.</td>
<td>Models are temporary in nature. A model is changed when its behavior is not in agreement with observations of the target.</td>
</tr>
<tr>
<td>Van Der Valk et al. (2007)</td>
<td></td>
<td></td>
<td></td>
<td>As part of the research activities, a model can evolve through an iterative process.</td>
</tr>
<tr>
<td>Source</td>
<td>Most naïve ideas observed</td>
<td>Low-level</td>
<td>Moderate</td>
<td>Expert or expert-like</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Grosslight et al. (1991)</td>
<td>Scientists could have different views of the same entity</td>
<td>Most naïve ideas observed</td>
<td>One could emphasize different aspects of the entity, omitting or highlighting certain things to provide greater clarity</td>
<td>A scientist could have more than one model for the same thing because different models could be used to address different specific interests or questions about the referent; there could be several competing models for the same thing</td>
</tr>
<tr>
<td>Schwarz and White (2005)</td>
<td>Multiple models could not exist</td>
<td>Multiple models could exist because different people conducted their experiments differently or because different people have different opinions or ideas about the phenomena</td>
<td>Different models arise because of aspects such as variability and error</td>
<td>Different people may have different interpretive frameworks for the same data</td>
</tr>
<tr>
<td>Grünkorn et al. (2014)</td>
<td>Only one final and correct model</td>
<td>Multiple models may exist because of differences between different model objects.</td>
<td>Multiple models may exist because the original allows the creation of different models.</td>
<td>Multiple models may exist because of different hypotheses about the original.</td>
</tr>
<tr>
<td>Krell and Krüger (2017)</td>
<td>Not discussed</td>
<td>Different models result from different modeler’s ideas OR from focusing on different aspects of the target</td>
<td>Different modeler’s ideas represent competing models or theories for explaining the target phenomenon</td>
<td>Different models for the same phenomenon result from different assumptions about the target or addressing different aspects of the target.</td>
</tr>
<tr>
<td>Crawford and Cullin (2005)</td>
<td>Different models are the result of different learning modalities, educational levels, audiences, or forms</td>
<td>Different models result from different modeler’s ideas OR from focusing on different aspects of the target</td>
<td>Different modeler’s ideas represent competing models or theories for explaining the target phenomenon</td>
<td>Different models for the same phenomenon result from different assumptions about the target or addressing different aspects of the target.</td>
</tr>
<tr>
<td>Van Der Valk et al. (2007)</td>
<td></td>
<td></td>
<td>Two or more different models can represent the same target, but in practice, there may be reason to select one model over others.</td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 3  Comparison of literature accounts of qualitative levels of knowledge about evaluation of models

<table>
<thead>
<tr>
<th>Source</th>
<th>Most naïve ideas observed</th>
<th>Low-level</th>
<th>Moderate</th>
<th>Expert or expert-like</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grosslight et al. (1991)</td>
<td>A model should be understandable.</td>
<td></td>
<td>A model should be exact, smaller, or proportional to the real thing</td>
<td>A primary guideline for making a model is to consider its purpose; purposes would be mediated by the extent of one's interest in structure, function, explanation, precision, predictive power, communication, and/or scope</td>
</tr>
<tr>
<td>Schwarz and White (2005)</td>
<td>All models have equal value</td>
<td>Models have equal value and there is no way to know which one is right</td>
<td>Some models are better than others due to their validity and accuracy</td>
<td></td>
</tr>
<tr>
<td>Grünkorn et al. (2014)</td>
<td>No testing of models</td>
<td>Models can be evaluated by testing the model object itself</td>
<td>Models can be evaluated by comparing the model with the original</td>
<td>Models can be evaluated by testing hypotheses about the original with the model</td>
</tr>
<tr>
<td>Krell and Krüger (2017)</td>
<td>Models are validated by the scientific community (an external authority).</td>
<td></td>
<td>Models are validated by comparing the behavior of the model with the behavior of the target.</td>
<td>Models can be checked or verified by comparing the results obtained by manipulating the model with the observations obtained in the real world.</td>
</tr>
<tr>
<td>Crawford and Cullin (2005)</td>
<td>Models are validated by the scientific community (an external authority).</td>
<td></td>
<td>Models are validated by comparing the behavior of the model with the behavior of the target.</td>
<td>Models can be checked or verified by comparing the results obtained by manipulating the model with the observations obtained in the real world.</td>
</tr>
<tr>
<td>Van Der Valk et al. (2007)</td>
<td>Not explicitly discussed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Process of modeling</td>
<td>Source</td>
<td>Most naïve ideas observed</td>
<td>Low-level</td>
<td>Moderate</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------</td>
<td>--------------------------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td></td>
<td>Grosslight et al. (1991)</td>
<td>Discussion centered around model criteria that are important to consider when designing or creating models</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Schwarz and White (2005)</td>
<td>Discussion centered around the process of changing models, which we consider to be better aligned with model changeability</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grünkorn et al. (2014)</td>
<td>Not discussed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Krell and Krüger (2017)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crawford and Cullin (2005)</td>
<td>Connection between modeler’s ideas and the model what the modeler thinks rather than what they are trying to get across.</td>
<td></td>
<td>Get the model to behave like the target. This would result in different relationships being built into the model.</td>
</tr>
<tr>
<td></td>
<td>Van Der Valk et al. (2007)</td>
<td></td>
<td></td>
<td>The construction of a model requires creativity, among others in finding a compromise between “having analogies with” and “being different from” the target, so as to optimally serve its purpose.</td>
</tr>
</tbody>
</table>
instruction; the lower anchor is often defined empirically (National Research Council, 2007). Developers of construct maps have taken different approaches, including empirical approaches, to defining benchmarks along the continuum of knowledge between the lower and upper anchors, termed the “messy middle” by Gotwals and Songer (2010). For instance, Berland and McNeill (2010) analyzed elementary, middle, and high school students’ classroom discourse and written work products to identify the middle levels of progression in students’ ability to construct scientific arguments, between complex, expert-like argumentation and very simple argumentation.

Much of the extant literature on construct mapping is closely related to literature on learning progressions, and there exists a variety of ways experts conceptualize the relationship between construct maps and learning progressions (Wilson, 2009a). For instance, according to Wilson (2009a), a set of related construct maps may together form a learning progression. To illustrate, Schwarz et al. (2009) developed a learning progression comprising two dimensions of metamodeling knowledge, which they described as a pair of related construct maps. Using the paired construct maps, the authors observed that students progressed from thinking about models primarily as illustrative tools to understanding models as explanatory tools and that their ability to revise and critique models improved.

We emphasize, however, that we do not conceptualize the four construct maps developed in this manuscript as constituting a learning progression and acknowledge Gilbert and Justi’s (2016) argument that a complete learning progression for models and modeling would be intertwined with other learning progressions such as for argumentation and visualization. We do, however, believe that the construct maps we have developed may be useful tools as the basis of learning progressions-oriented instructional design; The Framework for K-12 Science Education (National Research Council, 2012) has recently highlighted learning progressions as a promising approach for developing curricula which support increasing understanding of practices (such as modeling) across grade levels. We also see the construct maps presented here as potentially useful for design of robust, evidence-centered assessments. Construct modeling has been highlighted by a recent report by the National Research Council (2014) on developing assessments aligned with K-12 science standards. Furthermore, multiple choice assessments based on construct maps have been shown to be useful diagnostic tools (e.g., Briggs et al., 2006; Hadenfeldt, Bernholt, Liu, Neumann, & Parchmann, 2013).

In the next sections, we discuss our approach to the two steps of the analysis: developing hypothetical construct maps and refining these construct maps using survey data.

4 METHODS

Our development of construct maps as assessment tools focused on four dimensions: model changeability, multiplicity, evaluation, and the process of modeling. We chose these dimensions in part because we define scientific models as human-constructed tools used for explaining and predicting natural phenomena (Schwarz et al., 2016). As such, we view the constructs of the nature and purpose of models as deeply intertwined and inseparable from other constructs. For instance, models can be evaluated or changed according to how well they serve their intended purpose and there may be multiple models for a given phenomenon because they serve distinct purposes. Though the dimensions of nature and purpose of models clearly hold value for the science education research community (Crawford & Cullin, 2005; Grosslight et al., 1991; Grünkorn et al., 2014; Schwarz & White, 2005), we do not discuss them here in part because
these dimensions have been well-explored by other scholars (e.g., Gogolin & Krüger, 2018; Schwarz et al., 2009), and because we see them as intertwined with the four dimensions specified.

4.1 | Participants and setting

Participants for this study were students enrolled in a first-semester introductory chemistry course, the first in a two-semester sequence, at a research-intensive university in the Midwestern United States. Instructors delivered course material primarily through lecturing, and the text for the course was the 12th edition of *Chemistry: The central science* by T. L. Brown, Lemay Jr., Bursten, Murphy, and Woodward (2012). Students enrolled concurrently in discussion, laboratory, and case study sections of the course. Discussion sections were led by teaching assistants and were designed to provide problem-solving practice in an active-learning environment. Case studies prepared students for the laboratory activities and offered examples of real-world applications of chemistry concepts from the course.

4.2 | Development and administration of the models in chemistry survey

The focal survey items were part of the models in chemistry survey (MCS), a largely open-ended survey instrument developed by our research group. The pilot version of the assessment included 23 forced-choice and open-ended prompts based on existing literature regarding students’ metaknowledge of models and modeling (Schwarz & White, 2005; White, Collins, & Frederiksen, 2011). The final question asked students to indicate whether they would be willing to be contacted for participation in an additional interview. The items from the MCS we discuss in this study are shown in Table 5. The full survey is included in the supplemental information. Findings from remaining questions of the MCS, which focus on students’ ideas about characteristics of specific chemical models, are discussed in another manuscript (Lazenby, Rupp, Brandriet, Mauger-Sonnek, & Becker, 2019; Lazenby, Stricker, Brandriet, Rupp, & Becker, 2019).

We piloted the MCS with introductory chemistry students in the spring of 2017 and recruited eight respondents to participate in semi-structured interviews examining the response-process

<table>
<thead>
<tr>
<th>Item #</th>
<th>Construct</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Process of modeling</td>
<td>How do you think models are developed?</td>
</tr>
<tr>
<td>4</td>
<td>Evaluation of models</td>
<td>How do you think scientific models are evaluated (that is, how do you think scientists determine if a model is “good” or not?)</td>
</tr>
<tr>
<td>5</td>
<td>Model changeability</td>
<td>Once a scientific model is developed, do you think scientists would ever change it? (Yes/No)</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Why might scientists (not) change a scientific model?</td>
</tr>
<tr>
<td>7</td>
<td>Model multiplicity</td>
<td>Do you think two scientists could come up with two different models for the same real-world observation? (Yes/No)</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Why do you think two scientists (could or could not) come up with different models for the same real-world observation?</td>
</tr>
</tbody>
</table>
validity of survey items. We selected interviewees via a maximum variation sampling approach; specifically, we identified a sample of students whose survey responses reflected a range of sophistication. In the interviews, students were asked to describe their thought process as they worked through each question in the MCS. The interviewer (the third author) asked students to elaborate on how they understood the prompts and their reasoning about the prompts. Interviews were video recorded and students’ written work was captured using a Livescribe pen (Livescribe, 2017). Participants were compensated for their time with a $10 gift card.

In the analysis of data from the pilot survey and interviews, we focused on the alignment of students’ responses with the target construct of each question. In particular, we focused on identifying instances of either construct-irrelevance (i.e., measurement of constructs other than the target construct) or construct-underrepresentation (i.e., failure to measure the full range of the target construct) (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Based on the findings, we modified several MCS question prompts. The final version of the MCS includes 23 forced-choice and open-ended prompts as well as a series of demographic questions.

We administered the final version of the MCS via Qualtrics (2017) in the week before final exams in Fall 2017. The findings we report here, therefore, represent students’ epistemic knowledge of models and modeling just before completing a full semester of undergraduate introductory chemistry. Students received three extra credit points for completing the survey and had the option to either consent to or decline having their data analyzed for research. Extra credit points were awarded regardless of whether students consented to research participation. We collected 864 response sets to the MCS from the 1,017 students enrolled in the course. Twenty students submitted multiple sets of responses; only the first set of responses from these students was retained for analysis. The response sets of twelve students under the age of eighteen and seventy-nine students who declined to participate in the study were also removed from the data set. Finally, the responses of sixteen students who provided only very incomplete or off-construct answers were removed from the study. The final sample includes the response sets of 757 students, most of whom were 18–21 years old (93.9%) and in their first semester in college (79.6%). Students’ gender identification was representative of the class as a whole; 45.7% identified as male, 52.7% identified as female, less than 1% identified as non-binary, and 1% did not respond to the question. We obtained Institutional Review Board approval for all data collection.

4.3 | Data analysis

We used accounts of students’ knowledge of models and modeling in the extant literature (Tables 1–4) to develop initial deductive coding schemes to analyze the data collected via the MCS (Crawford & Cullin, 2005; Grosslight et al., 1991; Grünkorn et al., 2014; Krell et al., 2015;
We then used a constant comparative approach to refine the coding schemes, adding additional inductive codes as needed to capture any emergent themes (Glaser, 1965).

We developed four construct maps based on prior accounts of students' metamodeling knowledge (Crawford & Cullin, 2005; Grosslight et al., 1991; Grünkorn et al., 2014; Justi & Gilbert, 2002; Schwarz & White, 2005) that discussed qualitatively distinct ways in which students might think about model changeability and model multiplicity.

### TABLE 7
Construct maps for model changeability and model multiplicity and cross-construct themes identified at Levels 2 and 3

<table>
<thead>
<tr>
<th>Level 3</th>
<th>Model changeability</th>
<th>Model multiplicity</th>
<th>Cross-construct theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student response indicates that iterative change of models is part of typical research activity; responses in this category recognize the role of the modeler’s interpretation of data as the agent of model change</td>
<td>Student response indicates that multiple models may be the result of different interpretations of or explanations for the same data; includes responses which discuss complementary models or models created for different purposes</td>
<td>Interpretation of data is an important aspect of modeling; recognition of the role of the scientist/modeler</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 2</th>
<th>Model changeability</th>
<th>Model multiplicity</th>
<th>Cross-construct theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student response indicates that a model may be changed if new data or information about the target phenomenon are discovered; responses in this category do not discuss interpretation of data but rather discuss the data itself as the agent of model change</td>
<td>Student response indicates that multiple models are the result of focusing on different aspects of the target phenomenon or applying different methods to study the phenomenon</td>
<td>Data is important in modeling; recognition of the role of data in modeling, but data is the main focus, not the modeler</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Model changeability</th>
<th>Model multiplicity</th>
<th>Cross-construct theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student response indicates that a model may be changed if there is something wrong with them; response ultimately converges on the idea that there exists one correct model</td>
<td>Student response indicates that multiple models may exist due to differences in modalities, educational levels, disciplinary field, etc.; includes responses such as “everyone thinks differently”</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 0</th>
<th>Model changeability</th>
<th>Model multiplicity</th>
<th>Cross-construct theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student response indicates that models cannot or should not be changed</td>
<td>Student response indicates that multiple models may exist, but only one correct model may exist for a given phenomenon</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other</th>
<th>Model changeability</th>
<th>Model multiplicity</th>
<th>Cross-construct theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student response is too vague for interpretation or unrelated to the target construct</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
students think about models and modeling. We also used Van Der Valk et al.'s (2007) account of experts’ conceptions of models and modeling to define the upper anchors of the construct maps. When comparing literature-based construct maps to the focal data, we found that prior findings on students’ ideas about model changeability and multiplicity aligned well with our data. In contrast, prior findings on students' ideas about the evaluation of models and the process of modeling did not fully represent the focal data and thus we used an inductive approach to define some levels of the construct maps for these constructs. We describe this process in more detail later in the manuscript.

4.3.1 | Reliability

To assess the reliability of the coding structures, we conducted an inter-rater reliability study. For the construct maps related to model changeability, model multiplicity, evaluation of models, and the process of modeling, the first author acted as the primary coder and the second author independently coded approximately 20% of the responses to each task (150 randomly selected responses for each task). We calculated percent agreement and Cohen's kappa (Cohen, 1960), an index of inter-rater reliability. Because using construct maps for analysis suggests an ordered nature, we calculated two versions of Cohen's kappa (Table 6), one that assumes the codes are nominal (not ordered) (Cohen, 1960) and another that assumes ordered codes (Cohen, 1968). The nominal version penalizes all disagreements equally, while the ordinal version penalizes disagreements that are farther apart more severely. Cohen's kappa values exceeding 0.70 provide substantial evidence of reliability (Table 6) (McHugh, 2012).

5 | FINDINGS

5.1 | Model changeability and multiplicity

As noted earlier, we developed literature-based hypothetical construct maps and empirically validated the maps for each construct. The hypothetical construct maps for model changeability and model multiplicity accounted for the entire range of observed responses (very naïve [Level 0] to expert-like [Level 3]) for the tasks related to model changeability and model multiplicity. We identified themes across both constructs (which we refer to as cross-construct themes) for Levels 2 and 3 (Table 7); for example, for both model changeability and model multiplicity, the definition of Level 3 indicates that a student is aware of the interpretative role that a modeler or scientist plays in the interpretation of data. In the following sections, we describe the themes we observed and provide student quotes which exemplify the types of responses assigned to each theme.

5.1.1 | Model changeability

To elicit students' ideas about the changeability of scientific models, we first asked, “Once a scientific model is developed, do you think a scientist would ever change it?” (forced-choice; Yes, I think scientists might change a model/No, I don't think scientists would change a model). This forced-choice prompt was followed by an open-ended task in which students were asked either,
Why might a scientist change a scientific model? or Why might a scientist not change a scientific model? depending on their selection in the forced-choice task. An overwhelming majority of students (98%; n = 739 of N = 757) selected “Yes, I think scientists might change a model.” Despite the near unanimity of the responses to the forced-choice prompt, there were a range of responses to the open-ended task, spanning from the naïve ideas that models only change if they are incorrect or should not change at all to expert-like acknowledgement that models will inevitably change because they are human-constructed explanatory and predictive tools.

The most sophisticated responses (Level 3), which we consider to be expert-like responses based on their similarity to Van Der Valk et al.’s (2007) description of practicing scientists’ understanding of scientific models and modeling, explained that the iterative change of models is a normal part of research and may be due to either a modeler’s new interpretation of data or a new purpose for the model. For example, a typical response in this category stated, “There are always new discoveries and new discoveries lead to different ways of looking at things, so because perspectives change, so do scientific models.” This response shows that the student recognizes that ongoing scientific activity may lead scientists to revise their understanding of real-world phenomena, and in turn lead to model revision. For the construct of model changeability, 12% (n = 90) of responses were assigned to Level 3 (Figure 1).

We categorized 59% (n = 449) of responses on the changeability of models as Level 2 (moderate-level). Consistent with the cross-construct theme—the recognition that empirical data and experimentation play some role in modeling—these responses typically explained that new data or other evidence can lead to changes in models but did not discuss the interpretation of empirical evidence. Rather, Level 2 responses portrayed the data or evidence itself as the agent of model change. One typical response stated, “They might change a scientific model if they come across new evidence or research that allows for a new, improved model.” While this student noted that the discovery of “new evidence or research” allows scientists to improve an existing model, there was no mention of the role of the scientist as the interpreter of this new evidence.

Some students maintained that models change if they were previously “wrong” or “incorrect.” We assigned these responses to Level 1. Responses of this type are less sophisticated than Level 2 and Level 3 responses because the former potentially represent the incorrect belief that
models are either “correct” or “incorrect.” Additionally, responses of this type imply that a single “correct” model exists, when in fact, models are human-constructed tools that simplify or highlight certain aspects of a target phenomenon and can be useful or not useful, and multiple useful models can exist for a given phenomenon. The following explanation is typical of a Level 1 response: “If the previous model was proven wrong, the model can be modified to depict the truth.” The statement that a model can “depict the truth” indicates a fundamental misunderstanding of the nature of models as human-constructed tools. Furthermore, the reference to models being “proven wrong,” or the model’s ontological status, can be interpreted as an endorsement of the belief that a single “right” model exists. We assigned 26% (n = 195) of the responses to Level 1.

A small number of responses stated that models should never change; We classified these responses as Level 0. Students in this category typically explained that “incorrect” models should be replaced instead of changed. For example, one student concluded, “If you change a model then it means the previous model had its flaws and therefore a new model would have to take its place and the old one would be trashed.” We assigned only 2% of student responses to Level 0. The remaining 1% (n = 9) of responses were either too vague to interpret or did not focus on the focal construct. We categorized these responses as Other.

5.1.2 | Model multiplicity

To elicit students’ ideas about the multiplicity of scientific models, we first asked, “Do you think that two scientists could come up with two different models for the same real-world observation?” (forced-choice; Yes/No). This forced-choice prompt was followed by an open-ended task in which students were asked either, “Why do you think that two scientists could come up with different models for the same real-world observation?” or “Why do you think that two scientists could not come up with different models for the same real-world observation?” depending on their prior selection. As with the model changeability prompt, an overwhelming majority of students (98%; n = 739 of N = 757) selected “Yes” to the forced-choice prompt. However, analysis of students’ reasoning about why scientists could develop multiple models for a given observation revealed a range of ideas, spanning from naïve conceptions that while multiple models may exist, only one is correct, to expert-like notions that multiple models may be the result of alternative interpretations of empirical data or observations.

As noted earlier, there was a cross-construct theme in Level 3 responses, in which students recognized the role of the scientist/modeler in model development. We assigned responses that said that multiple models can result from different interpretations of or explanations for the same observations to Level 3. For example, one student explained, representative of responses assigned to this category, “Two scientists can come up with different models for the same observations because they both could have different interpretations of what the observation is. The way scientists interpret the data and observations affect how the models look compared to other scientists’ models.” Twenty percent (n = 149) of responses discussed the multiplicity of models in a manner consistent with Level 3.

Some students recognized that the same data and observations can yield multiple models for a given phenomenon but explained this outcome by noting that scientists might focus on different aspects of a single phenomenon (Level 2). For example, one Level 2 student concluded, “Two scientists could try to explain two different aspects of a real-world observation. This is common with highly complex topics like explaining how an atom works.” Other responses in
this category mentioned that scientists could use different methods to study the target phenomenon, yielding different models. A typical response of this type stated, “Two scientists may implement different experimental methods to explain one process that may result in different models.” In each of these responses, students recognized the role that data play in model development but did not discuss the role of the modeler as an interpreter of the data. These types of responses are consistent with Level 2 in the construct map; 17% (n = 131) of students discussed the multiplicity of models in a manner consistent with Level 2.

Nearly half of all responses (n = 343; 45%) attributed the existence of multiple models solely to differences in the modelers themselves, such as preferred learning modalities, educational levels, or disciplinary specialty, and did not mention the role of data. Representative of Level 1 responses, one student explained, for example, “Some scientists might develop a more visual model with pictures and some might prefer representing numbers. It depends on perspective and how a scientist chooses to represent information or an idea.” Another common theme was that, “Everyone thinks differently.” We consider this type of response indicative of a lower level of understanding (Level 1) than those described earlier (Levels 2 and 3) because the former do not provide any evidence that the student understands the role that empirical data/observations play in model development, instead attributing differences in models to differences in modelers.

The most naïve response type we observed (Level 0) were the answers that stated that multiple models may exist, but there is only one “correct” model for a given phenomenon; Students who selected “No” for the forced-choice prompt were also assigned to this category. A typical response in this category asserted, “If they come up with different models, it means one of them did it wrong since they were unable to achieve the same result.” We consider these responses the most naïve type because they provide no evidence of an understanding that models are human-constructed representations distinct from the target phenomenon. Thirteen percent (n = 101) of student responses were consistent with Level 0. The remaining responses were uninterpretable (too vague to categorize) or off-construct (e.g., provided an example of multiple models for a given phenomenon, such as multiple historical models of the atom). We classified these responses as Other (n = 34; 4%).

5.2 The evaluation of models and the process of modeling

The constructs of the evaluation of models and the process of modeling have received less attention in the literature than model changeability and multiplicity (Grosslight et al., 1991; Grünkorn et al., 2014; Krell et al., 2015; Schwarz & White, 2005; Van Der Valk et al., 2007). When we applied the deductive approach described earlier to build construct maps for the evaluation of models and the process of modeling, we found that the extant literature accounts did not accurately represent the full range of responses we observed in the data. Therefore, we used a combination of extrapolation of cross-construct themes and inductive analysis to build parallel construct maps for these two constructs that fully represented the themes in survey responses (Table 8).

We focused on extrapolating cross-construct themes to predict how students with moderate-to expert-level knowledge would discuss the evaluation of models and the process of modeling but did not use this approach extensively for students with very naïve or low-level ideas. This approach is theoretically reasonable because students who possess expert-like knowledge are likely to have coherent conceptions of models and modeling across constructs while those with lower-level knowledge likely have less coherent knowledge structures across constructs.
<table>
<thead>
<tr>
<th>Model changeability</th>
<th>Model multiplicity</th>
<th>Evaluation of models</th>
<th>Process of modeling</th>
<th>Cross-construct theme</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 3</strong> Student response indicates that iterative change of models is part of typical research activity; responses in this category recognize the role of the modeler's interpretation of data as the agent of model change.</td>
<td>Student response indicates that multiple models may be the result of different interpretations of or explanations for the same data; includes responses which discuss complementary models or models created for different purposes.</td>
<td>Student response indicates that models may be evaluated based on their ability to predict or explain real-world phenomena, empirical results may be compared to model-predicted results.</td>
<td>Student response indicates that models are developed based on scientists' interpretations of or explanations for data or experimental results; includes responses which discuss analyzing and finding patterns or trends in data.</td>
<td>Interpretation of data is an important aspect of modeling; recognition of the role of the scientist/modeler.</td>
</tr>
<tr>
<td><strong>Level 2</strong> Student response indicates that a model may be changed if new data or information about the target phenomenon are discovered; responses in this category do not discuss interpretation of data but rather discuss the data itself as the agent of model change.</td>
<td>Student response indicates that multiple models are the result of focusing on different aspects of the target phenomenon or applying different methods to study the phenomenon.</td>
<td>Student response indicates that models are evaluated by comparing the model with the target phenomenon; responses in this category include references to empirical testing or repeatability and the accuracy of the model compared to the target phenomenon.</td>
<td>Student response indicates that models are developed based on research and experimentation but do not discuss interpretation of empirical data.</td>
<td>Data is important in modeling; recognition of the role of data in modeling, but data is the main focus, not the modeler.</td>
</tr>
<tr>
<td><strong>Level 1</strong> Student response indicates that a model may be changed if there is something wrong with them; response ultimately converges on the idea that there exists one correct model.</td>
<td>Student response indicates that multiple models may exist due to differences in modelers; responses in this category attribute differences in models to differences in learning modalities, educational levels, disciplinary field,</td>
<td>Student response indicates that models are evaluated by their ability to communicate what the model represents, and be easily interpreted by all audiences.</td>
<td>Student response indicates that models are developed by scaling or copying the target phenomena to create a visual representation.</td>
<td>Scientific models are visual representations that are useful for teaching and learning about scientific phenomena.</td>
</tr>
<tr>
<td>Model changeability</td>
<td>Model multiplicity</td>
<td>Evaluation of models</td>
<td>Process of modeling</td>
<td>Cross-construct theme</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------</td>
<td>----------------------</td>
<td>---------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Level 0 Student response indicates that models cannot or should not be changed</td>
<td>Student response indicates that multiple models may exist, but only one correct model may exist for a given phenomenon</td>
<td>Student response indicates that models are evaluated by the scientific community (an external authority)</td>
<td>Student response indicates that models are developed through trial-and-error, educated guessing, or following a standard set of procedures such as “The Scientific Method.”</td>
<td>Level 0 responses represent students’ the most naïve ideas observed, which we consider to be fundamentally incorrect conceptions about the nature of scientific inquiry</td>
</tr>
<tr>
<td>Other</td>
<td>Student response is too vague for interpretation or unrelated to the target construct</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
To analyze responses related to the construct of the evaluation of models, we drew particularly from Crawford and Cullin’s (2005) account of the ways students’ might think about validating models. Mirroring the findings for the constructs of model changeability and multiplicity, we identified cross-construct themes for the top two levels (Levels III and IV in Crawford and Cullin’s study [2005] and Levels 2 and 3 in this manuscript). Because students who demonstrated very naïve knowledge about model evaluation expressed ideas that have not been reported in the extant literature, we defined Levels 0 and 1 based on an inductive analysis of the sample responses.

The construct of the process of designing and creating models has also received little attention in the literature; with the exception of students who demonstrated expert-like (Level 3) knowledge of this construct, the themes reported in the literature were not well-aligned with the themes observed in the data. Therefore, we defined Level 2 for this construct by extrapolating cross-construct themes, and we defined Levels 0 and 1 via inductive analysis of the sample responses.

5.2.1 Evaluation of models

To elicit students’ ideas about how to evaluate the quality of scientific models, we asked, “How do you think that scientific models are evaluated? (that is, how do you think scientists determine if a model is ‘good’ or not?)” As for the earlier items, we observed a range of responses from naïve to expert-like ideas about the evaluation of scientific models.

We identified responses that discussed models’ ability to explain and predict real-world phenomena, usually in terms of comparing model-generated data and empirical data, as expert-like responses (Level 3). As noted earlier, the definition for this category was based on the extant literature (Crawford & Cullin, 2005) and the extrapolation of the Level 3 cross-construct theme (recognition of the roles of both the modeler and empirical data). The following response is typical of Level 3: “I think models are evaluated based on their proximity to real-world observations and data and their ability to explain both past and future events.” Only 10% (n = 77) of responses to the evaluation of models prompt were assigned to Level 3.

Nearly half (48%; n = 361) of the responses to this prompt were assigned to a Level 2 (moderate-level) understanding of the evaluation of models. We classified responses as Level 2 if they discussed using empirical data to evaluate scientific models but did not mention the modeler’s role in this process (that is, comparing empirical data to the model). A typical Level 2 response noted, “A scientific model is evaluated by going through different experiments to determine if it [the model] does work.” Because the student did not elaborate on how a scientist might “determine if it [the model] does work,” the response was coded as Level 2.

An additional 12% (n = 92) of responses were assigned to Level 1. We coded responses that stated that models are evaluated based on their ability to communicate with or be interpreted by all audiences as Level 1. These responses are more naïve than Level 2 and 3 responses because they indicate that these students think of models as visual teaching tools rather than explanatory and predictive tools used by the scientific community. Responses of this type fail to recognize the role of empirical observation in modeling practice in any capacity. The following response is typical of Level 1: “[A model is good] if the model can be easily understood, is visually appealing.”
The most naïve responses (Level 0) were those that explained that models are evaluated by an external authority, such as other scientists. For example, one Level 0 response noted, “They rely on other scientists to check the model. The model has to be easily repeated with similar results. If something is wrong, hopefully, another scientist will notice.” We consider these responses more naïve than those in Levels 1 through 3 because they represent a potential misunderstanding of the scientific endeavor, including perhaps, the belief that the peer review process, or science in general, often includes replication of the experiments conducted by scientific colleagues, such as in the following example, “Peers of the creator will review and evaluate it based on their own findings.” The remaining 10% (n = 74) of responses were non-interpretable or off-construct and thus were assigned to the Other category.

5.2.2 The process of developing models

To elicit students’ ideas about the process of developing scientific models, we asked, “How do you think that scientific models are developed?” Like the previously described items, this open-ended prompt produced a wide range of responses.

Level 3 responses, which recognize the role of both the modeler and empirical data in the process of developing models, typically discussed scientists’ interpretation of data, sometimes including finding patterns or trends. A typical Level 3 respondent noted, “Scientific models are developed by a large amount of observation and data collection. Patterns observed among the data can lead to the development of a model.” Nine percent (n = 65) of the responses to this prompt were categorized as Level 3, which was defined by extrapolating the cross-construct theme identified for Level 3 for the previous constructs.

Level 2 responses recognized the role of data in the process of developing scientific models but did not describe how the data becomes a model. For example, one student stated, “Scientific models are developed through analysis of research and the information acquired through the research.” A large proportion (n = 300; 40%) of responses were categorized as Level 2, which was also defined via the extrapolation of cross-construct themes. Responses in this category recognized that empirical data collection plays a role in modeling but did not discuss how data might become a model.

Level 1 responses indicated that models are developed by scaling down or copying the target phenomenon to create a visual representation. We defined this level empirically using inductive analysis. Level 1 responses are less sophisticated than Level 2 and 3 responses because the former neither show that the student understands the role of empirical data in modeling nor acknowledge that models need not be visual representations. The following is a typical response from this group: “They [models] are developed by taking what something looks like or a simple version of it and making it into something you can see.” Only 5% (n = 35) of responses were categorized as Level 1.

Level 0 responses for this construct noted that models are created by simple “trial-and-error” or by following a predetermined set of steps, such as “the scientific method.” We defined Level 0 responses inductively and we consider these responses the most naïve because they represent a fundamental misunderstanding of how science progresses. We categorized 17% (n = 126) of the responses related to the process of modeling as Level 0.

We categorized the remaining 31% of responses to this prompt as Other. Of the 231 responses in this category, 53 (7% of total responses) focused on who develops models rather than the intended construct, while 73 (10% of total responses) addressed the purpose of developing models rather than the intended construct. Another ten responses addressed both who develops models and the purpose of modeling but not the intended construct.
The prevalence of responses that did not address the target construct may indicate an issue(s) with the item prompt. We address this possibility in the limitations section of the manuscript. The remaining 115 responses in the Other category were too vague for interpretation or did not address any related construct.

6 | LIMITATIONS

One potential limitation of the study is our assumption that students’ responses to the open-ended tasks fully represent their epistemic knowledge of the four focal model-related constructs. Students’ ability to articulate their ideas about models and modeling has been shown to be sensitive to context (Gobert et al., 2011; Krell et al., 2014; Moritz Krell et al., 2015; Lazenby, Rupp, et al., 2019; Lazenby, Stricker, et al., 2019). As such, students may have possessed knowledge that was not activated by the domain-general prompts and the written format.

However, the observed themes from the survey data generally mirror those described in both the extant literature and the think-aloud interviews which we used to provide evidence of response-process validity. For example, for the item asking about the process of modeling, many students discussed related constructs such as who might engage in modeling or the purpose of modeling. As in the surveys, we also observed a number of responses that did not mention any related construct (total off-construct responses = 31%). Although none of the students who participated in the response-process validity interviews expressed difficulty interpreting this prompt, the prevalence of off-construct responses may suggest limited or fragmented metamodeling knowledge, or that participants may not have understood the question as intended. We believe the former is likely as students’ responses are reflective of their curricular experiences; because students in this traditional chemistry course had limited opportunities to engaged in the process of modeling, it is not surprising that there was variation in the extent to which students were able to discuss the process of modeling.

A second limitation of the study is the use of a cross-sectional approach to data collection in a traditional course in which students who had not received explicit instruction on models or modeling, in contrast to a longitudinal approach to assessing the development of students’ metamodeling ideas in a modeling-rich environment (Eylon & Linn, 1988). It is thus possible that the ideas elicited from these students are not representative of all students. Future research should seek to validate the construct maps for use in assessing student growth over time and for other student populations.

Finally, we acknowledge that the construct-modeling approach we have utilized makes some important assumptions about how learners’ knowledge progresses. Sikorski and Hammer (2010) caution against the consideration of more expert-like or canonical ideas as “more sophisticated” as well as the idea that progress can be modeled as a series of successive levels. Likewise, (Duncan & Rivet, 2013) caution that intermediate level ideas, such as ideas in the “messy middle” (Gotwals & Songer, 2010), may not be accurately modeled by linear progressions such as construct maps or learning progressions.

7 | CONCLUSIONS

In this study, we describe four empirically validated construct maps that address first-year undergraduate chemistry students’ epistemological knowledge about model changeability,
model multiplicity, the evaluation of models, and the process of modeling. Our findings regarding development of two metamodeling constructs (model changeability and multiplicity) are consistent with prior literature findings about how students’ ideas progress (Crawford & Cullin, 2005; Grosslight et al., 1991; Grünkorn et al., 2014; Moritz Krell & Krüger, 2017; Schwarz & White, 2005). We also elaborate progressions for model evaluation and the process of modeling that reflect ideas not previously reported in the literature.

We report two cross-construct themes that we believe represent critical transitions in development of students’ metamodeling knowledge. The cross-construct theme for Level 2 represents a shift towards recognition of the role of empirical data plays in modeling; however, at this level students are still unable to demonstrate understanding of the modeler’s interpretative role in making sense of data. The Level 3 cross-construct theme relates to recognition of the interpretative role that a modeler or scientist plays in the analysis of data, which we believe represents a shift towards more sophisticated thinking about models and modeling.

In our analysis of general chemistry students’ ideas about models and modeling, we found for the three constructs—model changeability, the evaluation of models, and the process of modeling—student responses classified as Level 2 were the most prevalent. This distribution of responses suggests that many general chemistry students may be “getting stuck” at Level 2 and will likely require additional opportunities to progress towards understanding the role of modelers in interpretation of data.

For the remaining construct, that of model multiplicity, Level 1 responses were the most prevalent. Students’ whose responses were classified as Level 1 attributed the existence of multiple models to differences in modelers (e.g., different representational preferences), rather than differences in choices about how systems are modeled based on data. Here too, we believe students may need opportunities to recognize the role of data and data interpretation in the construction of multiple models for phenomena.

With regard to both model multiplicity and model changeability, many students discussed the existence of various historical models of the atom, a topic covered early in the course and one of the only times that the nature of models was discussed during the semester. These observations are not particularly surprising given that students primarily engage with scientific models in formal educational settings, and students’ perceptions of models and modeling are highly reflective of their instructional experiences.

Across constructs, we observed that students often discussed models as teaching tools or visual representations that are useful for learning, rather than as scientific tools useful for predicting and explaining phenomena. For example, with respect to model evaluation, 12% of responses mentioned that models should be easy to understand and interpret for non-scientist populations (e.g., students). This too may be understandable given that students may have seen scientific models primarily in formal learning environment.

Finally, for all constructs, those students who expressed a Level 0 understanding seemed to activate intuitive and sometimes fragmented and off-construct ideas about aspects of the scientific endeavor, consistent with a Knowledge-in-Pieces perspective (diSessa, 1988, 2018). For example, some thought that models are either “right” or “wrong” rather than useful or not useful while others stated that science proceeds through pure trial-and-error. Again, these ideas likely reflect students’ instructional experiences, such as using trial-and-error or “the scientific method” in grade school science. For these students, too, the opportunity to engage in thinking about models may improve their metamodeling knowledge.
8 | IMPLICATIONS

Models are the explanatory and predictive tools which are key in making sense of the natural world for practicing scientists, and as such, it is important for students to develop expertise in developing and using models as they prepare to be scientists (Schwarz et al., 2016). Even for students who will not join the scientific community, the ability to consider the nature and purpose of models is an important component of scientific literacy (Berland et al., 2016; Greene & Yu, 2015; Schwarz et al., 2016).

The current results suggest that students have some useful intuitive ideas about models and modeling that may be refined and built upon by instruction. For example, many participants understood that models can be evaluated according to their communicative ability with a variety of audiences (evaluation of models, Level 1; 20%). Fewer students, however, considered the role of data in the evaluative process. If given the opportunity, however, students might integrate this additional knowledge element into their existing knowledge systems (diSessa, 1988, 2018) regarding models and modeling.

To help their students fully understand models and modeling, instructors must not only teach students to use models to solve problems, but must also offer opportunities for students to create, evaluate, and revise models. One opportunity to support students’ ideas engagement with models and modeling in chemistry is through use of modeling-focused curricular resources that have been developed for undergraduate chemistry classes (Tien et al., 1999), and laboratory settings (Csizmar et al., 2013; Edwards & Head, 2016; Tien et al., 1999; Wolfson, Hall, & Branham, 1996). We believe that explicit discussions about epistemic ideas, for instance, what counts as a model, when accompanied with engagement in modeling practice, may support students’ epistemic ideas about models and modeling. Opportunities to think about models, as well as with models, will allow students to reorganize their existing knowledge about models and modeling and integrate new knowledge into existing structures, thereby helping them develop richer metamodeling knowledge.

As we have noted, there is a critical need for assessments that enable researchers and educators to assess growth in students’ ideas about models and modeling. To date, evidence of the impact of modeling-focused curricular resources at the undergraduate level, where available, has largely focused on their effect on students’ ideas about the particulate nature of matter (Tien et al., 1999; Tien, Teichert, & Rickey, 2007). Efforts to assess students’ epistemic ideas about models and modeling have largely used selected-response assessments such as the SUMS instrument to assess students’ perceived metamodeling knowledge (e.g., Burgin et al., 2018; Cheng et al., 2014; Everett et al., 2009; Gobert et al., 2011; Levy & Wilensky, 2009; Park et al., 2017; Treagust et al., 2002). Since SUMS is a Likert-scale self-report assessment, it has limited utility in monitoring growth in metamodeling knowledge.

Tools such as construct maps are therefore important for allowing researchers and instructors to assess growth in students’ knowledge and ability from a developmental perspective (National Research Council, 2014; Wilson, 2009a, 2009b). As such construct maps such as those presented here may be useful as a basis of assessment design (e.g., Briggs et al., 2006; Hadenfeldt et al., 2013) or curricular reform (e.g., Berland & McNeill, 2010; Schwarz et al., 2009). Prior work has focused primarily on use of construct maps for development of metamodeling knowledge of K-12 students (Schwarz & White, 2005); in this study, we have shown that our construct maps can be a useful tool for valid and reliable assessment of undergraduate STEM students as well.
We suggest that the construct maps presented here support design of assessments that can be used to examine growth in students’ ideas about models and modeling. While as we have noted, construct maps are distinct from scoring guides, we believe the construct maps presented here could serve as a basis for scoring guides or rubrics which could be used in conjunction with open-ended assessment items designed to probe students’ ideas about models and modeling. While all survey items discussed in this manuscript examined students’ domain-general metamodeling knowledge, we argue that the resulting construct maps may be useful assessment tools in either domain-specific or domain-general contexts.

The construct maps may also serve as a basis for the design of selected-response items addressing metamodeling knowledge. Our own ongoing work focuses on development of ordered multiple choice assessments based on these construct maps that can be used by instructors as diagnostic tools (e.g., Briggs et al., 2006; Hadenfeldt et al., 2013). Our hope is that the maps can help instructors meet students where they are at and use curricular resources to make the best use of students’ existing ideas.

ACKNOWLEDGMENTS
This material is based upon work supported by the National Science Foundation under grant DUE-1611622 and the National Science Foundation Graduate Research Fellowship Program under Grant No. 1650114. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

ORCID
Katherine Lazenby https://orcid.org/0000-0002-9672-8631

REFERENCES


**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

---

**How to cite this article:** Lazenby K, Stricker A, Brandriet A, Rupp CA, Mauger-Sonnek K, Becker NM. Mapping undergraduate chemistry students’ epistemic ideas about models and modeling. *J Res Sci Teach*. 2019;1–31. [https://doi.org/10.1002/tea.21614](https://doi.org/10.1002/tea.21614)