1-4-2021

From Ideas to Items: A Primer on the Development of Ordered Multiple-Choice Items for Investigating the Progression of Learning in Higher Education STEM

Katherine Lazenby
University of Iowa

Morgan E. Balabanoff
University of Nebraska - Lincoln

Nicole M. Becker
University of Iowa

Alena Moon
University of Nebraska - Lincoln

Jack Barbera
Portland State University, jbarbera@pdx.edu

Citation Details

Published as Katherine Lazenby, Morgan E. Balabanoff, Nicole M. Becker, Alena Moon, and Jack Barbera, "From Ideas to Items: A Primer on the Development of Ordered Multiple-Choice Items for Investigating the Progression of Learning in Higher Education STEM," Journal of Chemical Education, DOI: 10.1021/acs.jchemed.0c01121.

From Ideas to Items: A Primer on the Development of Ordered Multiple-Choice Items for Investigating the Progression of Learning in Higher Education STEM

Katherine Lazenby[a], Morgan E. Balabanoff[b], Nicole M. Becker[a], Alena Moon[b], and Jack Barbera[c*]

[a]Department of Chemistry, University of Iowa

[b]Department of Chemistry, University of Nebraska - Lincoln

[c]Department of Chemistry, Portland State University

ABSTRACT

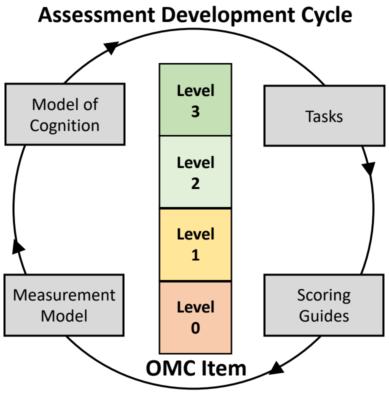
Identifying effective methods of assessment and developing robust assessments are key areas of research in chemistry education. This research is needed to evaluate instructional innovations and curricular reform. In this primer, we advocate for the use of a type of assessment, ordered multiple-choice (OMC), across postsecondary chemistry. OMC assessments are grounded in a developmental perspective, which treats students' knowledge as developing in sophistication over time. This is in contrast to a dichotomous perspective, which asserts students' knowledge is either aligned or misaligned with scientifically accepted knowledge. By drawing on a developmental perspective, OMC assessments offer insights into student understanding that can be useful for informing instruction. To that end, this primer will overview OMC assessments, illustrate their development and evaluation in two chemistry contexts, and make an argument for their utility in the chemistry education community.

KEYWORDS

General Public, Chemical Education Research, Testing/Assessment

Reference Information:
Katherine Lazenby, Morgan E. Balabanoff, Nicole M. Becker, Alena Moon, and Jack Barbera, "From Ideas to Items: A Primer on the Development of Ordered Multiple-Choice Items for Investigating the Progression of Learning in Higher Education STEM," *Journal of Chemical Education*, DOI: 10.1021/acs.jchemed.0c01121.

GRAPHICAL ABSTRACT



INTRODUCTION

"*If science teaching is about facilitating student learning, then we need to find out whether students have learnt. For that we need to find out what students know or understand, both before and after teaching, to judge if there has been any learning.*" (Taber[1], p. 3)

As eloquently stated by Taber in this quote, assessment is at the heart of teaching. It is how educators and education researchers gather insights about learners and the impacts of educational practices. The burden on educators to gather this insight is necessary because of how students learn. They build their own understanding through their learning experiences.[2] The premise that learners build their own understanding based on prior knowledge and experiences sets up the basis for a rather complex trajectory of learning. Knowledge becomes more "expert-like," forming more coherent and useful knowledge structures, as the learner engages in learning experiences.[3]

The development of expert-like knowledge structures is informed by ideas of conceptual change. The most prominent operationalization of conceptual change in chemistry education is the idea of learners holding "misconceptions," which are stable conceptions that do not align with scientific knowledge.[4,5] Inherent in this view is a dichotomous perspective of learner conceptions; that is, they either align with scientifically normative knowledge or they do not. Early perspectives of conceptual change sought to correct these misconceptions by establishing cognitive dissonance through exposing the learner to the inadequacy of their conceptions.[6] More recent perspectives have sought to catalogue non-normative ideas held by students, which can then be combated in instruction. This cataloging has been achieved through the development of a large body of diagnostic instruments called concept inventories. The strength of a concept inventory lies within each individual item's capacity to uncover the non-normative fragments of learners' knowledge of that topic. However, this cataloging approach has been challenged by Cooper and Stowe[3] who conclude that "*misconceptions are not necessarily well-characterized monolithic entities to be readily and formulaically surmounted.*" Therefore, a concept inventory may generate information about which misconceptions a learner finds plausible before and/or after instruction. However, assessment grounded in a dichotomous perspective can fail to capture the complexity of learning posited by constructivism.

In contrast to this dichotomous perspective is the idea that learners' understanding progresses over time to become more sophisticated,[7] where sophisticated knowledge is generally viewed as a stable, cohesive, productive, and flexible network of knowledge elements.[8,9] This idea has resulted in learning progressions, which have been defined as "*successively more complex ways of thinking about an idea that might reasonably follow one another in a student's learning.*"[10] This positions all learners somewhere in their development of productive and complex knowledge. Framing student understanding with this developmental perspective moves educators away from a deficit model of what students do not know to one of dynamic development and reweaving of ideas.[3] It is important to clarify here that learning progressions do not imply that all learners follow the same developmental steps, or in the same order; rather, they represent possible learning trajectories.[10] Even so, they offer useful models for situating and interpreting student performance.

A developmental perspective has implications for assessment. Beyond cataloging incorrect conceptions, assessments should offer information about the structures of students' knowledge that can be situated in models about how those structures morph through learning. For these reasons, learning progressions have been advocated for as a means of aligning curriculum and assessment[11,12] and have been used to track student progress.[13] While early work on learning progressions in science education dates back nearly 15 years,[10,14] there has been a resurgence of interest as compiled in a recent themed issue of *Chemistry Education Research and Practice*.[15] For educators to be able to harness the information provided by a learning progression, assessments must be developed and aligned with the progression.

While Holme et al.[16] call on the CER community to create assessments aligned with developmental theories of learning, the extant CER literature only contains a few assessments that are explicitly linked to developmental perspectives on learning.[17–21] Hadenfeldt et al.[18] developed an Ordered Multiple-Choice (OMC) assessment, a type of assessment to be described in detail below, to measure secondary students' understanding of the structure and composition of matter. Similarly, Scalise and colleagues[17] created the Perspectives of Chemists framework and characterized university-level chemistry students' development of understanding of matter, energy, and change based on constructed-response assessments; students' responses were scored according to the developmental progression. Building on this work, Wei et al.[19] administered an assessment comprising both multiple-choice and constructed-response items to assess students' understanding of matter, which were similarly scored based on the Perspectives of Chemists framework. Park et al.[20] developed OMC assessment items to assess high school students' learning related to matter and energy and the use of models in chemistry. Such assessments have been shown to be useful for making valid and reliable inferences about student knowledge levels as well as for identifying instructional strategies that leverage students' existing 'stepping stone' knowledge and abilities. Two examples of reform-oriented curricular approaches for general chemistry, Chemical Thinking, and Chemistry, Life, the Universe and Everything (CLUE), are explicitly linked to evidence-based learning progressions in the ways that chemists think and practice; likewise, assessments in both curricula are linked to the learning progressions that frame the course.[22,23] Here, we discuss one approach to developing assessments aligned with developmental theories of learning.

Ordered multiple-choice (OMC) assessments

One approach to assessment that is aligned with a developmental perspective on learning and learning progressions uses Ordered Multiple-Choice (OMC) items. Within an OMC item, each response option is explicitly linked to a developmental level from a construct map (Figure 1), which is a representation of different "levels" of sophistication along a continuum of knowledge for a specific progress variable.[7] The key feature of a construct map is that for a progress variable, there be qualitatively different performances that can be organized hierarchically according to sophistication.[7,24] For this reason, assessments comprised of a set of OMC items can provide information about learners' performance on a construct of interest well beyond that of traditional (dichotomously scored) multiple-choice assessments.[25]

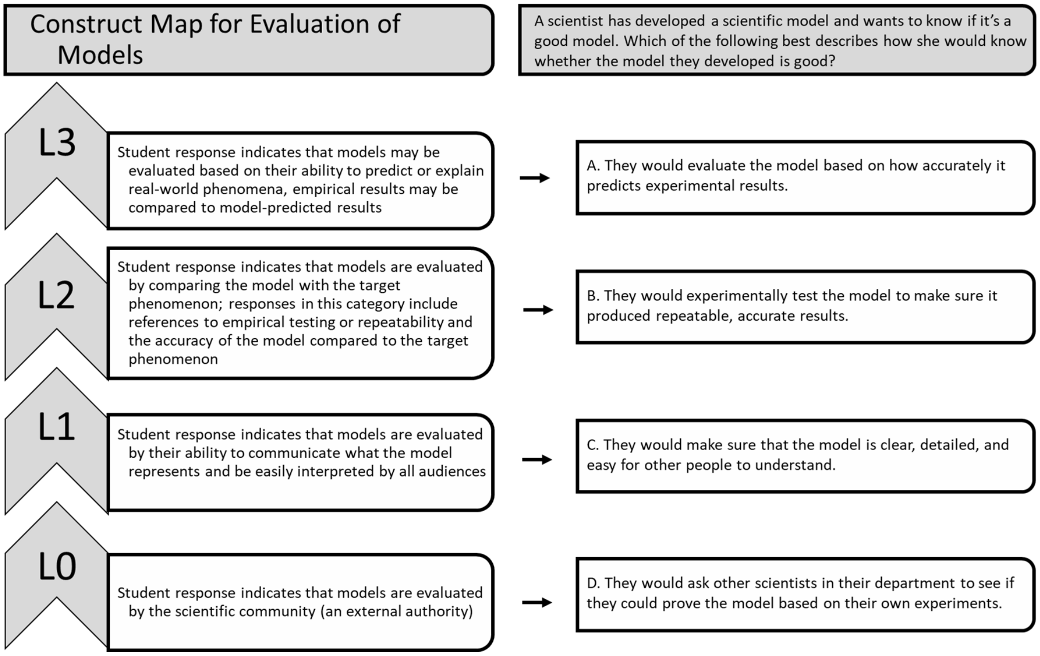

Construct Map for Evaluation of Models
A scientist has developed a scientific model and wants to know if it's a good model. Which of the following best describes how she would know whether the model they developed is good?
L3
Student response indicates that models may be evaluated based on their ability to predict or explain real-world phenomena, empirical results may be compared to model-predicted results
A. They would evaluate the model based on how accurately it predicts experimental results.
L2
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
B. They would experimentally test the model to make sure it produced repeatable, accurate results.
L1
Student response indicates that models are evaluated by their ability to communicate what the model represents and be easily interpreted by all audiences
C. They would make sure that the model is clear, detailed, and easy for other people to understand.
L0
Student response indicates that models are evaluated by the scientific community (an external authority)
D. They would ask other scientists in their department to see if they could prove the model based on their own experiments.

Figure 1. A construct map and an example of an OMC item based on the construct map. Each response option (A, B, C, D) is derived from a level of the construct map (L3, L2, L1, L0).

OMC assessments present a number of advantages over other assessment approaches. In comparison to open-ended assessments, OMC assessments can be scored more quickly, and at lower cost, and provide more unambiguous and precise indicators of student thinking. Additionally, Briggs et al.[25] argue that OMC assessments offer reliability advantages over open-ended assessments and have at least comparable, if not superior, potential for generating valid and reliable data compared to traditional multiple-choice assessments. Hadenfeldt et al.[18] demonstrated that OMC assessments can provide rich, diagnostic information about learners' levels of understanding equally as well as open-ended assessments. Such diagnostic information, from a set of OMC items based on the same progression (construct map), may support inferences about student understanding at the classroom or individual level and track learners' progress over time.[25]

Goal of this primer on OMC assessments

Given the movements toward reformed curricula and calls for assessments based on a developmental perspective of learning, we propose OMC assessments as a path forward for the

Chemistry Education Research (CER) community. Therefore, this primer introduces an assessment framework that emphasizes strong connections to a model of cognition. After the framework is explained and detailed, it is employed to outline the steps in developing OMC items for two different concepts within chemistry, namely, metamodeling knowledge and wave-particle duality. Therefore, this primer also serves as a companion piece to the full OMC instrument manuscripts that will result from these works. Much like the various protocols used when developing concept inventories,[26] our goal is to provide readers with an evidence-based assessment framework and relevant examples on which they can build. To do so, this primer is guided by the following questions:

1. What is the assessment model that informs OMC assessments?
2. How are OMC assessments developed for use in chemistry contexts?
3. For what purposes are OMC assessments useful in CER?

QUESTION 1: WHAT IS THE ASSESSMENT MODEL THAT INFORMS OMC ASSESSMENTS?

Within educational settings, inferences about student learning are derived from assessments, a process referred to as "reasoning from evidence."[27–29] This process was first presented as a triad known as the "assessment triangle" (Figure 2A)[30] which depicts the three key elements of an assessment as: 1) a model of student cognition, 2) a set of observations, and 3) a process of interpretation. It was reported that "for an assessment to be effective, the three elements must be in synchrony."[30] However, both before and after the introduction of the assessment triangle, it has been noted that most educational measures lack the *model of cognition* element, thereby limiting the interpretation of assessment scores.[27,31] To bolster the importance of the model of cognition, Brown et al.[31] characterize the practice of assessment as a four-step cyclical process (Figure 2B). Each step in this cycle is informed by the construct modeling approach such that the four "cornerstones" mediate and structure each step, resulting in a "*coherent system of assessment that is consistent with and expands upon the assessment triangle.*"[32] This "cornerstone" model is an expansion of the assessment triangle and has been proposed for use during the construction and evaluation of assessments to aid in the development of more robust measures, including those aligned with learning progressions.[31,32] This model was implemented by its developers when designing the Evidence-Based Reasoning Assessment

System[32] and the Measure of Conceptual Complexity[31,33] as well as within chemistry during development of the Performance assessment of Undergraduate Research Experiences (PURE) instrument.[34]

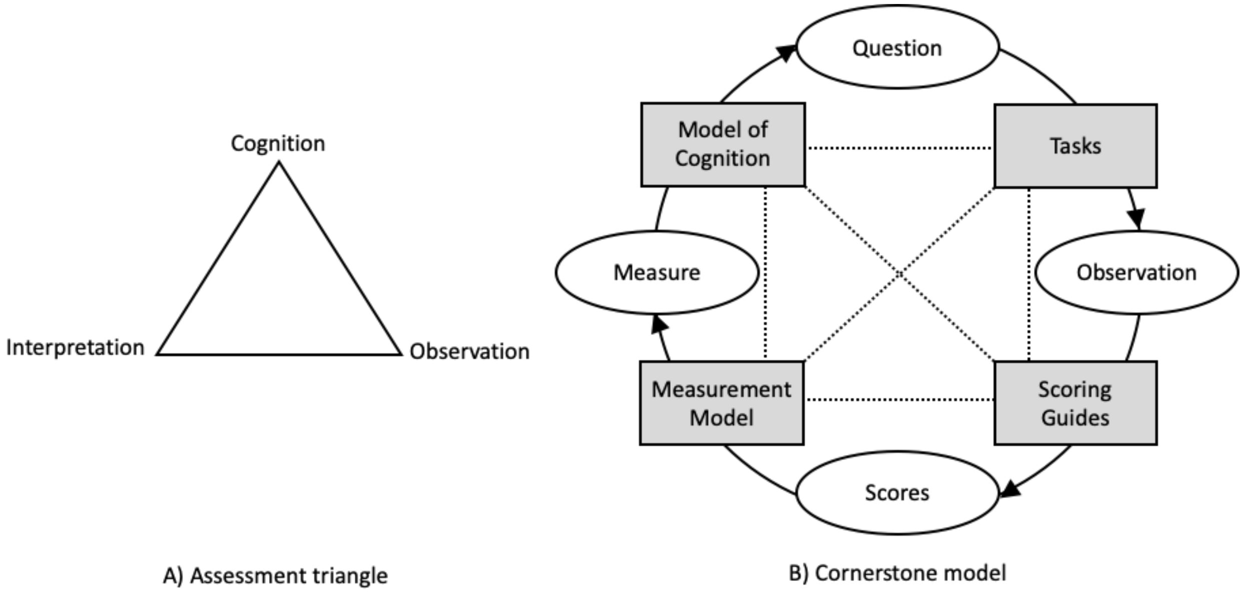


Figure 2. A) The assessment triangle, reprinted with permission from ref. 30. B) The cornerstone model, reprinted with permission from ref. 31. Steps of the assessment cycle appear in ovals and the cornerstones of assessment appear in blocks. Dashed lines represent the connections among assessment elements.

The cornerstone model depicts the four steps in an assessment cycle (ovals in Figure 2B) overlaid by the cornerstones of assessment (blocks in Figure 2B). In this model, an assessment is defined by all four cornerstones. The assessment cycle begins with a question about student learning. For example, based on a model of student cognition, one could ask “after instruction in [course], what are students’ understandings of [focal concept]?” From this start, the cycle proceeds to collecting observations through a set of designed tasks around the focal concept. The observations are transformed into scores through the application of a scoring guide. The scores are then turned into a measure of the focal concept by applying a measurement model to the data. The outcomes of this [focal concept] measure are then interpreted based on the model of cognition to answer the question. A salient feature of the cornerstone model is the interconnections among assessment elements (noted by the dashed lines in Figure 2B). In the next section, we explain each cornerstone and how they are operationalized within the construct modeling approach.[24]

Overview of each cornerstone

Model of Cognition

A model of cognition has been defined as a "*theory or set of beliefs about how students represent knowledge and develop competence in a subject domain.*"[30] Without a model of cognition, claims about assessment results are limited to characterizations of students' overall proficiency on something. With a model of cognition, however, we can make claims about the nature of the proficiency and how it develops, which are key to engaging and supporting learners.[31] By situating the learner's response in a model that provides information about how knowledge or competency of that domain develops, the assessment goes beyond a one-time snapshot of a learner's overall proficiency. That is, a model of cognition offers a frame for interpreting assessment results in a way that can powerfully impact instruction.[27,30]

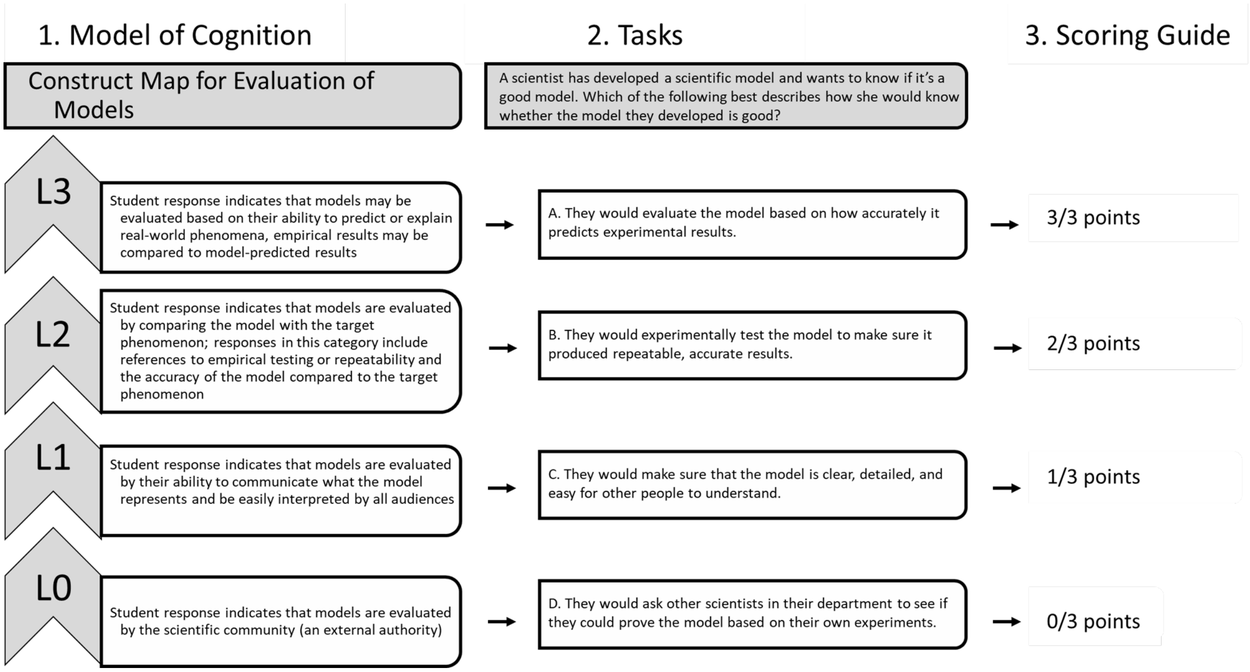
A key affordance of the OMC approach to assessment is its reliance on a robust model of cognition. The model of cognition that underpins OMC assessments is operationalized by a construct map (see example in Figure 1).[7,25,31] Construct maps are grounded in a developmental perspective and consequently acknowledge that learners vary in sophistication and demonstrate some level of sophistication with their assessment responses. Whereas, according to a dichotomous perspective, learners can be correct by being aligned with scientifically normative ideas or incorrect by holding any misaligned ideas. Learners demonstrate whether they are correct or incorrect, aligned or misaligned, with their assessment responses.[25] The model of cognition that aligns with this developmental perspective therefore models variation in the sophistication for a construct as a continuum from naïve intuition to expertise through a construct map, which transforms the continuum into different "levels" of sophistication for a progress variable. These different levels then directly inform OMC item development *and* interpretation of assessment responses.[7,25,31]

Construct maps can be generated through a combination of exploratory qualitative methods, existing theory and research findings, in addition to instructional expertise (i.e., knowledge of what represents expertise in a domain).[7] From any of these starting points, phenomenographic investigation of students' knowledge or competency in a domain results in qualitatively different categories, which can be organized hierarchically according to sophistication. The levels of a construct map make explicit the varying ways a learner can represent knowledge or competency for a progress variable (Figure 1). This product may appear to be very similar to a learning progression because in some ways it is. Learning progressions model varying ways a learner can represent knowledge for a big idea over a long period of time. A construct map captures the variation for a specific progress variable, which is why it is well suited to inform assessment development. Wilson[7] describes "*several different ways to see the relationship between the idea of a construct map and the idea of a [learning] progression*," the simplest and most straightforward being "*the learning progression as composed of a set of construct maps, each comprising a 'dimension' of the learning progression, and where the levels of the construct maps relate (in some way) to the levels of the learning progression.*"

Tasks and scoring guides

To determine students' levels of proficiency on a progress variable, tasks (i.e., items) need to be developed. The development of tasks must be preceded by a number of decisions about item type (e.g., open-ended or forced-choice) and descriptive characteristics (e.g., items are designed to be interpreted according to a specified model of cognition). For OMC items, the varying levels of a construct map directly inform the item response options, which means that a learner's response can be interpreted to place them within the construct map, making them "*likely to be useful, informative, and easily interpretable with respect to the latent proficiency.*" As such, diagnostic feedback is available to all responders in contrast to a dichotomous perspective, which only offers feedback to the learners with "incorrect" responses.[7,25,31]

Given the nature of OMC items, the tasks (cornerstone 2) and the scoring guide (cornerstone 3) are explicitly linked to each other (Figure 3), as well as to the model of cognition - represented by the construct map (cornerstone 1). In general, a scoring guide quantifies or orders the different kinds of student responses that a task will elicit. OMC items are therefore scored according to the development level of the construct map to which the response option is aligned. For example, in Figure 3, a student selecting answer option C, aligned with a Level 1 response on the construct map, would be assigned one out of three possible points. Once a set of items have been developed using this construct modeling approach, administration will result in a set of scores for each item and for each test-taker, which will be evaluated using a measurement model.[24,31]



A scientist has developed a scientific model and wants to know if it's a good model. Which of the following best describes how she would know whether the model they developed is good?

L3: 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 → A. They would evaluate the model based on how accurately it predicts experimental results. → 3/3 points

L2: 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 → B. They would experimentally test the model to make sure it produced repeatable, accurate results. → 2/3 points

L1: Student response indicates that models are evaluated by their ability to communicate what the model represents and be easily interpreted by all audiences → C. They would make sure that the model is clear, detailed, and easy for other people to understand. → 1/3 points

L0: Student response indicates that models are evaluated by the scientific community (an external authority) → D. They would ask other scientists in their department to see if they could prove the model based on their own experiments. → 0/3 points

Figure 3. An example of the first three "cornerstones" in OMC design (1. Model of Cognition [a construct map], 2. Tasks [an example OMC item, based on the construct map], and 3. Scoring Guide [assignment of numerical values for each answer option]). For simplicity, response options are displayed from highest to lowest level. However, when administered, response options are randomized.

Measurement model

The fourth cornerstone in construct modeling, the measurement model, relates a set of scores back to the construct map. The Rasch model family has demonstrated utility for evaluating the 'ordering' of OMC item responses and mapping their connections back to a construct map.[25,31] Rasch (and more broadly, Item Response Theory) models the relations between items and respondents probabilistically. Respondents are assumed to possess some amount (θ) of the latent trait of interest, which influences the likelihood that a respondent will select a specific answer option (forced-choice items) or provide a specific type of response (constructed-response items).[35,36] In the case of OMC items, respondents with higher θ values are more likely to select answer options which are aligned with higher levels on the construct map from which the item was derived.

Data analysis using Rasch yields estimates of item difficulty (δ) and person parameter (θ) that can be placed on the same logit (log-odds) scale. Estimating item difficulty and person parameter values on the same scale allows for the direct comparison of persons and items. Colloquially, Rasch estimates of the person parameter (θ) values are described as *ability* estimates, though θ can refer to estimates of respondents' aptitude, achievement, attitudes, personality traits, etc. We use the term *ability* in the colloquial way throughout our discussion of Rasch modeling and emphasize that OMC assessments need not measure ability as it is technically defined elsewhere.[37] Masters[38] adapted the dichotomous Rasch model for use with polytomous items with ordered scoring. This new model, called the partial credit model (PCM), has been used for modeling OMC items.[19,20,25] Masters[38] defines the PCM (Equation 1) as the probability that a person with ability value θ will score x on item i, where δ_{ix} is the item step difficulty parameter for the transition to the kth response option over the (k – 1)th and scores range from 0 to m (i.e., 0, 1, … m).

$$P_{ix}(\theta)=\frac{\exp[\sum_{x=0}^{m}\theta\text{-}\delta_{ix}]}{\sum_{x=0}^{m_i}\exp[\sum_{k=0}^{x}\theta\text{-}\delta_{ix}]} \qquad (1)$$

For ease of interpretation, it is common to convert difficulty estimates (δ_i) into Thurstonian threshold values; for polytomously scored items (such as OMC items), Thurstonian thresholds serve as indicators of "score difficulty."[39] For an OMC item with m answer options, m-1 Thurstonian thresholds may be calculated, which represent the ability value (θ) at which a respondent becomes more likely to select a higher-level response option over a lower-level response option.[35] For example, the first threshold (δ_{i1}) would be the ability value (θ) for selecting the Level 1 response option over the Level 0 option (Figure 3). These threshold values can also be thought of as performance levels at which a person becomes more likely to select the kth option over the (k – 1)th; each "step" results in a higher score on the item.[37] Figure 4 shows that a four-level item, like that shown in Figure 3, will result in three thresholds.

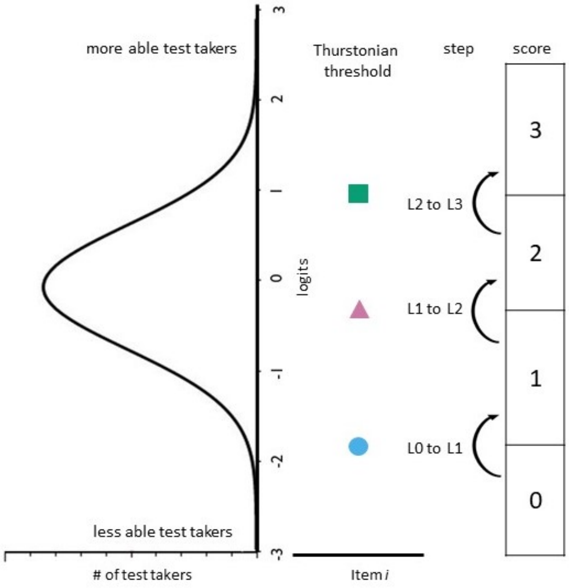


Figure 4. A 'step' interpretation of a hypothetical polytomous item with four ordered response categories (0, 1, 2, and 3) where colored shapes represent Thurstonion threshold values for item *i*. Threshold values represent the ability value (θ), in logits, at which students make the 'step' from one level (k-1th) to the next (kth).

A visual representation of the relations between respondents' ability value (θ) and difficulty parameters (δ_{ix}) on the logit scale is known as a Wright map (Figure 5). Wright maps can be used to evaluate if items are appropriately targeted (i.e., difficult) for the target population and, in the case of OMC items, to determine whether answer options reflect developmental progress (i.e., are 'ordered'), as expected.[25,35,38] Of the exemplar items shown in Figure 5, Item 2 is the most difficult item, indicated by the higher vertical position of each of the Thurstonian thresholds (colored shapes on right side of logit scale) compared to Items 1 and 3. In contrast, Item 1 is the easiest item. For each item, the response option difficulties are ordered as expected (i.e., the threshold value for moving to L1 is less than that for L2 which is less than that for L3). Finally, it can be determined that these three items are appropriately tailored to the target population, indicated by the overlap of the range in person abilities (histogram on left side of logit scale) and item thresholds.

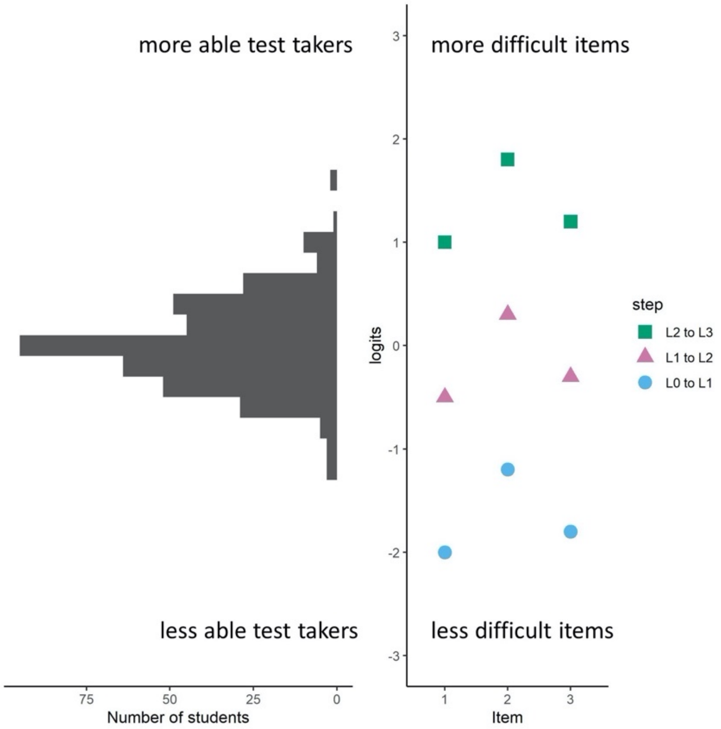


Figure 5. Example of a Wright map generated for three OMC items (labeled 1, 2, 3). Person ability distribution is shown to the left of the logit scale and item difficulty (Thurstonian thresholds) distribution is shown to the right.

An added feature of Rasch modeling, beyond placing person and item details on the same scale for direct comparison, is evaluation of the fit between the data from a set of items and the Rasch model. Mean square residuals (MNSQ) and standardized mean square residuals (ZSTD; the normalized t-score of the residual) are generated from an analysis and indicate the extent to which the data fit the Rasch model. It is typical to compute both INFIT (inlier-weighted MNSQs) and OUTFIT (unweighted MNSQs) for items; items with MNSQs (both INFIT and OUTFIT) between 0.7 and 1.3 and ZTSDs (both INFIT and OUTFIT) between -2 and 2 are considered to have good model-data fit.[40] When the PCM is used, these fit statistics extend to each Thurstonian threshold of an item. Therefore, they can be used when evaluating the quality of each item response level and provide added support for the construct map from which they were derived.

When incorporated into the cornerstone model, Rasch analysis and the resulting Wright maps for OMC items are useful for relating scores (from scoring guide; cornerstone 3) directly back to the construct map (model of cognition; cornerstone 1). Additionally, Rasch provides information about whether item 'step' difficulties are indeed ordered as expected, based on the progression in the construct map and whether the answer options for the tasks appropriately reflect the response patterns of the target population.

QUESTION 2: HOW ARE OMC ASSESSMENTS DEVELOPED FOR USE IN CHEMISTRY CONTEXTS?

Two different projects will be used to illustrate the development of OMC assessments in a chemistry context. OMCs are especially appropriate for chemistry contexts given the vast amount of work done on identifying and elucidating "big ideas" in chemistry. For example, the particulate nature of matter,[18,41–43] structure-property relationships,[44–46] and energy[47–50] are all contexts which have been studied extensively and on which students' thinking can vary in sophistication. The two projects that will be presented herein center around similarly big ideas in chemistry. The first project centers on students' knowledge about models and modeling, which will be referred to as *metamodeling knowledge*; the second focuses on students' understanding of the wave-particle duality of light, which will be referred to as *wave-particle duality*. In the remainder of this section, we will use examples from one or both projects to illustrate each cornerstone (Figure 2B) and to present our development processes.

Model of Cognition

Metamodeling

To develop construct maps related to undergraduate students' epistemic knowledge about models and modeling, we drew on both extant literature and students' responses to constructed response assessment tasks (details are provided in a recent manuscript by Lazenby et al.[51]). We developed four parallel construct maps using this approach – maps for Model Multiplicity, Model Changeability, Evaluation of Models, and The Process of Modeling. We determined that for each progress variable (construct map), prior literature and empirical data suggested four qualitatively distinct "levels" (Levels 0, 1, 2, and 3). We also identified themes across the four progressions, which we termed "cross-construct themes"; for instance, we identified that at Level 2, for all four constructs, students discussed the role of data in modeling phenomena but did not discuss the role of data interpretation. We conceptualize the relation between the four construct maps and the larger progression, which is represented by the cross-construct themes, similarly to Gotwals' description of nested progressions.[52] In this case, the construct maps represent a narrower focus on, for instance Evaluation of Models, which is nested in a larger, overall progression with a broader focus on metamodeling knowledge. For simplicity, here we focus only on one of the construct maps as a representation of the progression of knowledge related to Evaluation of Models and two OMC items derived from this construct map.

To explore students' conceptions of models, we designed and administered the Models in Chemistry Survey (MCS), which is comprised primarily of constructed-response items about the epistemic nature of scientific models and modeling.[51] The MCS was administered to first-semester university chemistry students near the end of the semester. We focus here on student responses to the MCS item prompt, "How do you think scientific models are evaluated (that is, how do you think scientists determine if a model is 'good' or not)?"

In building a construct map for students' ideas about the evaluation of models (Figure 6), we began by developing literature-based hypothetical construct maps from prior studies which characterized the qualitatively different ways that students and experts think about each of the specified constructs. For the construct of Evaluation of Models, we found that prior literature accounts did not represent the full range of responses observed in our population. To refine the literature-based construct map and develop an empirically informed construct map for general chemistry students' ideas about model evaluation, we used inductive analysis of students' MCS responses to identify emergent themes constituting progressively more sophisticated ideas. The levels of each construct map were iteratively revised in order to represent students' responses to the MCS items.

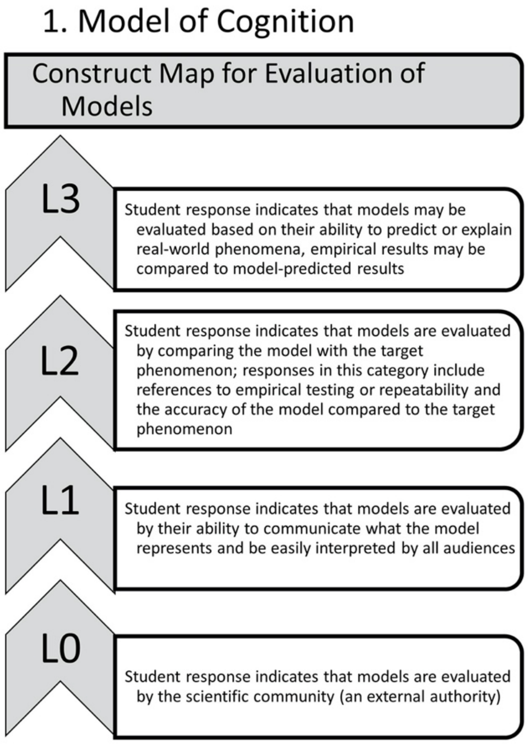

1. Model of Cognition
Construct Map for Evaluation of Models
L3
Student response indicates that models may be evaluated based on their ability to predict or explain real-world phenomena, empirical results may be compared to model-predicted results
L2
Student response indicates that models are evaluated 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
L1
Student response indicates that models are evaluated by their ability to communicate what the model represents and be easily interpreted by all audiences
L0
Student response indicates that models are evaluated by the scientific community (an external authority)

Figure 6. Empirically informed construct map for Evaluation of Models

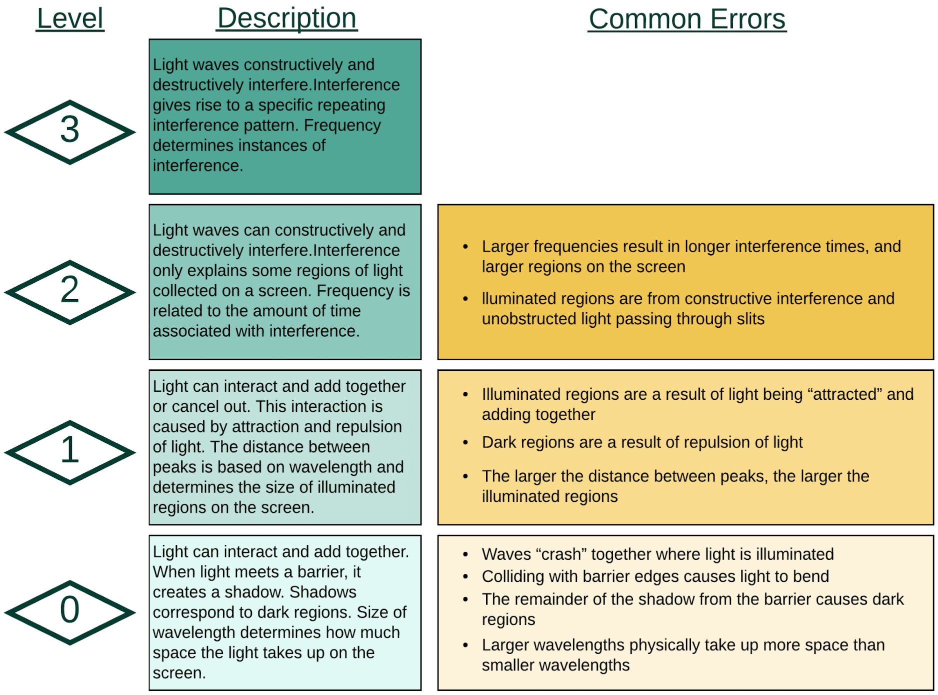
Wave-Particle Duality

The objective of our work was to model and assess students' understanding of the wave-particle duality of light. Targeting students' understanding of the dual nature of light is part of a larger objective to model chemistry student understanding of light-matter interactions. Wave-particle duality is an inherently quantum mechanical concept, which prompted us to grapple with students' movement between classical and quantum ideas.

Given the small amount of prior research available to draw on, in this project we began the construct map development process by conducting two cross-sectional qualitative studies using semi-structured interviews with chemistry students ranging in level (general chemistry to physical chemistry). One set of interviews centered around the photoelectric effect, which demonstrated particulate behavior of light, and required students to predict and explain the phenomenon using static representations. The second set of interviews centered around the double-slit experiment, which demonstrated wave behavior of light, and similarly required students to predict and explain. With the photoelectric effect, students engaged with static representations of the phenomenon, while with the double-slit experiment, students engaged with dynamic simulations. Both phenomena required students to draw conclusions about the behavior of light from experimental observations.

It became evident that explaining the complex phenomena of light-matter interactions presented several distinct challenges for students, even if the student successfully predicted an outcome. In particular we noted variation around students' ideas about the particulate behavior of light, wave behavior in the double slit context, wave behavior of light, and wave particle duality. To better capture the variation in students' reasoning, we opted to deconstruct the broader concept of light-matter interactions into four progress variables: Particulate Behavior of Light, Wave Behavior in Double-Slit Context, Wave Behavior of Light, and Wave-Particle Duality.

Our phenomenographic analysis of the interview data resulted in qualitatively different ideas that we organized hierarchically into construct maps for each of the four progress variables. A developmental perspective shaped the analysis by focusing our attention away from characterizing students' ideas as correct or incorrect, but rather more or less productive in explaining the phenomenon. More productive ideas were characterized by both alignment and span; that is, they were invoked and used appropriately across multiple contexts whereas less productive ideas were context dependent. In each construct map, the more productive ideas were considered to be higher levels of sophistication and less productive ideas constituted the lower levels of the construct map. Below is the Wave Behavior in Double-Slit Context construct map (Figure 7). Each level provides a description of student understanding and common errors. The common errors help clarify differences between levels and are resolved in the next level of the construct map.[7,53]



Level	Description	Common Errors
3	Light waves constructively and destructively interfere.Interference gives rise to a specific repeating interference pattern. Frequency determines instances of interference.	
2	Light waves can constructively and destructively interfere.Interference only explains some regions of light collected on a screen. Frequency is related to the amount of time associated with interference.	• Larger frequencies result in longer interference times, and larger regions on the screen • lluminated regions are from constructive interference and unobstructed light passing through slits
1	Light can interact and add together or cancel out. This interaction is caused by attraction and repulsion of light. The distance between peaks is based on wavelength and determines the size of illuminated regions on the screen.	• Illuminated regions are a result of light being "attracted" and adding together • Dark regions are a result of repulsion of light • The larger the distance between peaks, the larger the illuminated regions
0	Light can interact and add together. When light meets a barrier, it creates a shadow. Shadows correspond to dark regions. Size of wavelength determines how much space the light takes up on the screen.	• Waves "crash" together where light is illuminated • Colliding with barrier edges causes light to bend • The remainder of the shadow from the barrier causes dark regions • Larger wavelengths physically take up more space than smaller wavelengths

Figure 7. Construct map for Wave Behavior in Double-Slit Context progress variable

We envision a learning progression of light-matter interactions as being partially comprised of all four construct maps, where three are staggered in such a way that levels of one construct map are offset with regard to those of another construct map. Evidence suggests that the Particle Behavior of Light, Wave Behavior of Light, and Wave Behavior in Double-Slit Context construct maps are staggered, in that order, with increasing conceptual difficulty (or higher learning progression level). The fourth, Wave-Particle Duality, construct map would parallel and span the entire range of the three staggered maps. This is because having an understanding of the dual nature of light is not entirely dependent on a students' understanding of either wave or particle behavior. Rather, having the highest level of understanding of wave-particle duality is dependent on a student recognizing that the dual nature of light depends on observation and experimental setup. These staggered and parallel approaches to how related construct maps contribute to an overall learning progression are similar to those noted by Wilson[7] in his seminal work on the measurement of learning progressions and will be fully explicated in our full manuscript on this progression.

Summary

For the development of construct maps, we draw attention to 1) similarities between the two projects that serve as critical components of OMC development, 2) similarities that are coincidental, and 3) differences that are due to unique features of each project. In both projects, a combination of research literature and thematic analysis of qualitative data served to reveal knowledge varying in sophistication, which ultimately resulted in a construct map. This transformation of a wide range of knowledge representations within a domain, elicited through open-ended investigation, is a critical component of the model of cognition cornerstone. In both projects, the exemplar construct map consisted of four levels; this is a coincidence and construct maps may include any number of levels. There were some key differences between the projects, which source from the different starting points. Given a larger body of literature available to inform the metamodeling project, we were able to elicit student ideas through a constructed response survey. However, for the wave-particle duality project, we began by carrying out a very in-depth qualitative inquiry due to the small body of literature available. The starting point for the development of OMCs will depend largely on context and what is known about that context.

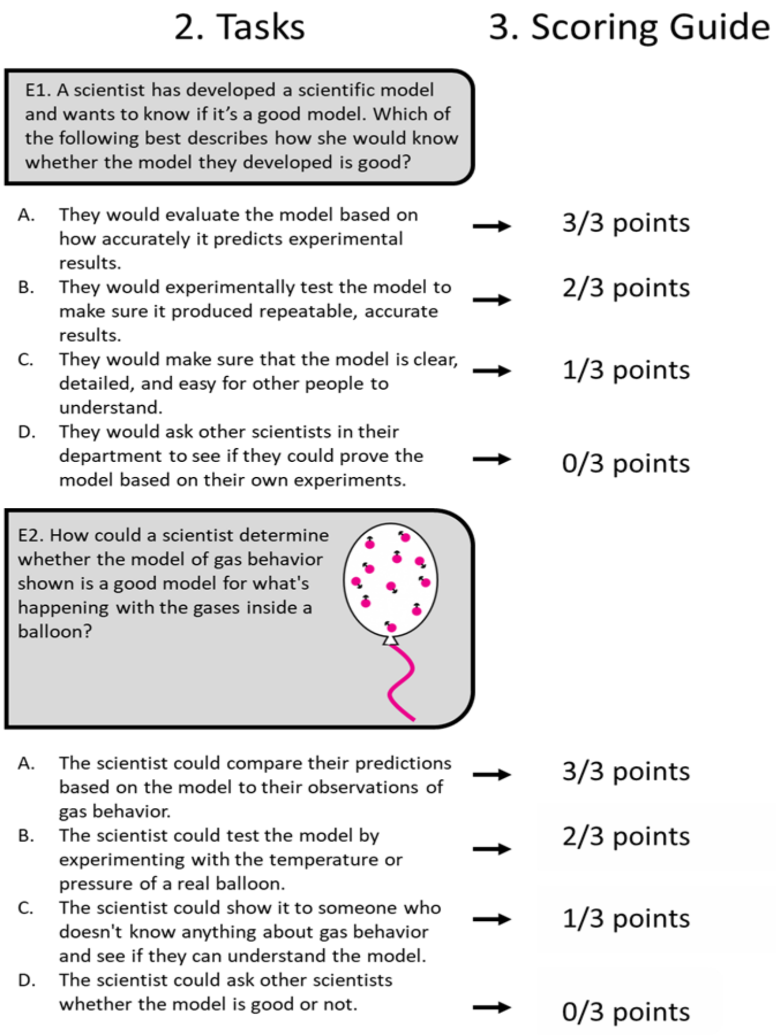
Tasks and Scoring Guide

As discussed earlier, the tasks and scoring guide cornerstones are inherently and explicitly linked for OMC assessments; therefore, they are discussed here together.

Metamodeling

Based on the four construct maps, built with regard to metamodeling ideas, we developed a set of pilot OMC items related to each construct (Evaluation of Models, Model Multiplicity, Model Changeability, and Process of Modeling) as well as items in both domain-general and chemistry-specific contexts. The domain-general items were written to parallel the constructed response items (from the MCS) used to develop the construct maps. For instance, the constructed response item from the MCS “How do you think scientific models are evaluated (that is, how do you think scientists determine if a model is “good” or not?)” was reframed as a forced-choice item “A scientist has developed a scientific model and wants to know if it’s a good model. Which of the following best describes how she would know whether the model they developed is good?” (Item E1, Figure 8).

Similarly, the response options on the domain-general items were written to reflect the types of constructed responses that were categorized into each level during our analysis of the MCS data. To do so, we identified exemplar responses for each level, which were used to create the response options to the OMC items. For example, one student's constructed response to the MCS prompt related to model evaluation stated, "Show it to other scientists, people who aren't scientists, or maybe show it to a science class and see if the students understand the model"; this response and others like it informed the writing of the Level 1 response option on the OMC instrument. This exemplar response was paraphrased into the OMC item response, "They would make sure that the model is clear, detailed, and easy for other people to understand."



2. Tasks

3. Scoring Guide

E1. A scientist has developed a scientific model and wants to know if it's a good model. Which of the following best describes how she would know whether the model they developed is good?

A. They would evaluate the model based on how accurately it predicts experimental results. → 3/3 points

B. They would experimentally test the model to make sure it produced repeatable, accurate results. → 2/3 points

C. They would make sure that the model is clear, detailed, and easy for other people to understand. → 1/3 points

D. They would ask other scientists in their department to see if they could prove the model based on their own experiments. → 0/3 points

E2. How could a scientist determine whether the model of gas behavior shown is a good model for what's happening with the gases inside a balloon?

A. The scientist could compare their predictions based on the model to their observations of gas behavior. → 3/3 points

B. The scientist could test the model by experimenting with the temperature or pressure of a real balloon. → 2/3 points

C. The scientist could show it to someone who doesn't know anything about gas behavior and see if they can understand the model. → 1/3 points

D. The scientist could ask other scientists whether the model is good or not. → 0/3 points

Figure 8. Two OMC items, E1 and E2, with corresponding scoring guides. For simplicity, response options are displayed from highest to lowest level. However, when administered, response options are randomized.

To produce chemistry-specific items, we identified models that are commonly presented in the first semester introductory chemistry course. This process was informed by course materials from previous offerings of the course, the course textbook,[54] and the Anchoring Concepts Content Map (ACCM).[55] For instance, we identified equations and representations related to gas behavior as candidate models which could be used as a context to probe students' chemistry-specific metamodeling ideas. Based on the identified model contexts, we developed an item bank of potential items and corresponding answer options that were aligned with the construct map for each item. For example, kinetic molecular theory was identified as a candidate model, as it is commonly presented in general chemistry courses. In these courses, gas behavior and kinetic molecular theory are commonly discussed in the context of expandable containers, like balloons; therefore, we constructed a representation of gas molecules inside of a balloon to situate item E2 (Figure 8). Response options for the chemistry-specific items are similar to those in the domain-general items; that is, they are derived from the construct maps and student language from the MCS data, though situated in the chemistry-specific contexts of models from the course.

Each of the OMC items was developed with three to four answer options, as appropriate. Each of the response options is aligned to a different level of the construct map; this is not a requirement for OMC items but rather a design consideration for this project. Because the answer options for each OMC item were derived from a specific level on a construct map (Levels 0 – 3), the items are scored (cornerstone 3) according to the level on the map they correspond to (e.g., selection of a Level 3 response receives 3 out of 3 possible points, while a selection of a Level 1 response receives only 1 of 3 possible points) (Figure 8).

For both domain-general and chemistry-specific items, the item writing process included multiple iterations. Items were iteratively revised based on 1) feedback from a team of chemistry education researchers, 2) response process interviews with general chemistry students, and 3) response data from a pilot survey, which included open-ended analogue versions of all of the items followed by the OMC items. We emphasize that the items themselves (and therefore the scoring guides), as well as the construct maps from which the items are derived, were revised iteratively based on multiple rounds of qualitative data collection and analysis (as represented by the dashed lines in Figure 2B). We further address item revision, based on quantitative analysis of response data, in our discussion of the fourth cornerstone (measurement model).

Wave-Particle Duality

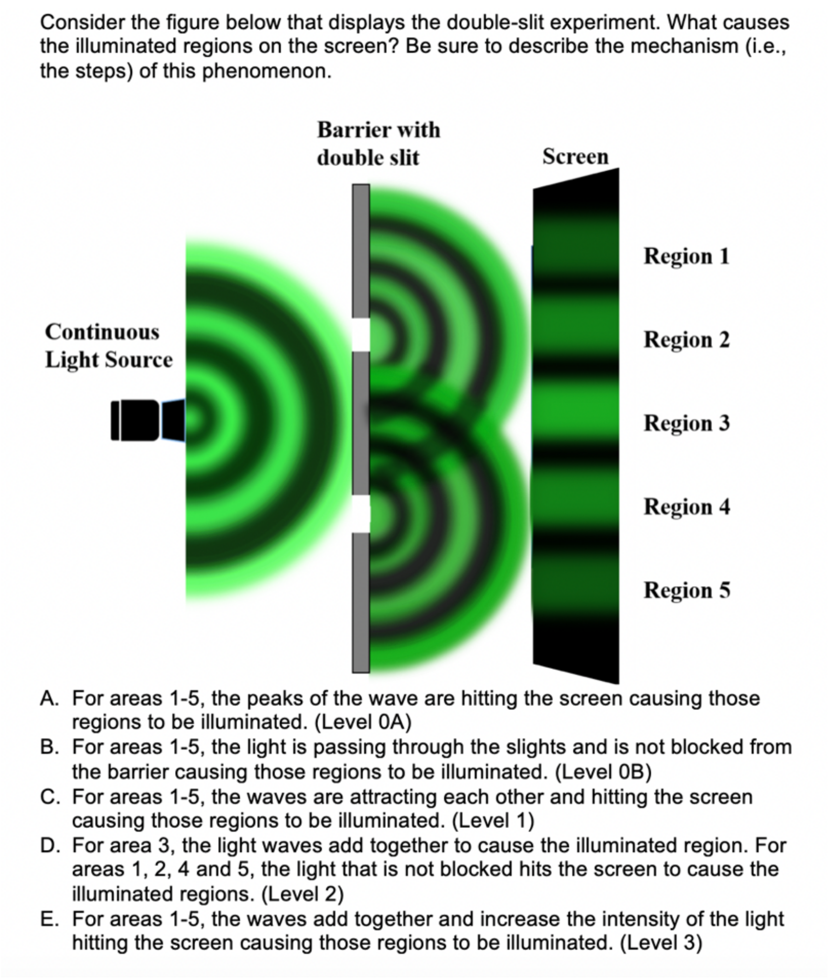
As in the metamodeling project, the construct maps for each of the four progress variables shaped the development of pilot OMC items. The OMC item stems paralleled our interview guides and similarly prompted students to make predictions and construct explanations. That is, the items were specifically designed to elicit the variation in student reasoning observed in the interviews. For example, tasks in the double-slit experiment interviews prompted students to explain various simulations. One simulation depicted a continuous light source and a barrier with two slits, similar to Figure 9. Students were then asked to explain the pattern on the screen. Therefore, as shown in Figure 9, the corresponding OMC item stem parallels the interview by asking students to provide explanations for the illuminated regions. Similarly, some item stems reflect the prediction questions posed in the interviews. For example, in the context of the photoelectric effect, interviews prompted students to predict the effect of increasing the intensity of a light above the threshold frequency. Again, a parallel OMC item stem was designed using similar language to elicit predictions.

As noted in the Model of Cognition section, two sets of interviews were conducted, one on the photoelectric effect and the other on the double-slit experiment. Analysis of the photoelectric effect interviews yielded particulate behavior items (n = 7), which targeted the particle behavior of light progress variable. The double-slit experiment interviews yielded two sets of items, double-slit experiment items (n = 4), which targeted the wave behavior in the double-slit context progress variable, and wave properties items (n = 3), which targeted the wave properties of light progress variable. Analysis of both interviews yielded wave-particle duality items (n = 4), which targeted the wave-particle duality progress variable. These items were designed to capture the range of student conceptions outlined in each progress variable's construct map.

Students' explanations in interviews were analyzed to generate item responses that captured student language. For example, in response to the Figure 9 prompt one student explained that, the only region that would illuminate due to constructive interference would be Region 3. *"It makes sense* *that of course we get two light sources on that screen from each of the openings. But then… in the middle [the light is] kind of combining… So the light that's coming out from the openings and it's going straight forward, it's unobstructed and it's not really interacting with anything else because it's just on its own straight forward path."* This type of explanation was seen with multiple students and was used to generate response option D in Figure 9. Student explanations were analyzed for similar themes and arranged hierarchically from least productive to most productive. Students' explanations were then used to generate the close-ended responses for each item.

Within each item, the levels from the appropriate construct map were used to further shape response options. Alignment with a construct map allowed for consistent language to be used across item responses and thoughtful consideration of the progress variables. Additionally, the identified common errors associated with each level helped provide specific and plausible response options. Figure 9 illustrates one OMC item targeting the Wave Behavior in Double-Slit Context progress variable, which is aligned with the construct map shown in Figure 7.

The pilot wave-particle duality OMC items included between three and five response options. Each response option for an item did not necessarily map to a unique level of a construct map. For example, in some items, a level may have been omitted if a unique response option representing that level was not noted in our interview data. In addition, a single map level may manifest as multiple response options to an item, as is the case for options A and B in Figure 9. The scoring guide for this item, as shown in Figure 9, would take the form: A=B=0, C=1, D=2, and E=3. That is, a response option earns the number of points that match the construct map level it aligns with. These pilot OMC items and their response options will be further revised based on additional student responses to both closed and open-response administrations, and further analysis of interview data.



Consider the figure below that displays the double-slit experiment. What causes the illuminated regions on the screen? Be sure to describe the mechanism (i.e., the steps) of this phenomenon.

A. For areas 1-5, the peaks of the wave are hitting the screen causing those regions to be illuminated. (Level 0A)
B. For areas 1-5, the light is passing through the slights and is not blocked from the barrier causing those regions to be illuminated. (Level 0B)
C. For areas 1-5, the waves are attracting each other and hitting the screen causing those regions to be illuminated. (Level 1)
D. For area 3, the light waves add together to cause the illuminated region. For areas 1, 2, 4 and 5, the light that is not blocked hits the screen to cause the illuminated regions. (Level 2)
E. For areas 1-5, the waves add together and increase the intensity of the light hitting the screen causing those regions to be illuminated. (Level 3)

Figure 9. An OMC item aligned with the Wave Behavior in Double-Slit Context construct map. For simplicity, response options are displayed from lowest to highest level. However, when administered, response options are randomized.

Summary

As with the development of construct maps in the Model of Cognition section, there are critical similarities and dissimilarities between the development of Tasks for the two exemplar content areas. Three critical components of OMC item design are demonstrated with both projects. The first component is that item stems were generated from qualitative investigations with students from the target population. This process helps to ensure that an item stem elicits a variety of student thinking for the domain of interest. The second component is that item responses map onto construct map levels, so that the construct map can be used to understand and interpret student responses. The third component is not uniquely critical to OMC item development, but rather is essential to the development of high-quality assessments in general. Items and item responses are iteratively revised based on student responses. Finally, the scoring guides took similar forms for each project, this is due to the fact that there are simply a limited number of ways in which to score OMC items.

Measurement Model

Metamodeling

Rasch analysis using the Partial Credit Model (PCM), and the resulting Wright Map, provides information about the item response step difficulties (δ_{ix}) (i.e., Thurstonian thresholds) in relation to both person ability values (θ) (horizontal interpretation) and the relative step difficulties of each response option for an OMC item (vertical interpretation).[38] For interpretation of this data to be meaningful, it is important to consider fit to the Rasch model and whether related items measure a single latent trait (i.e., that they are unidimensional), which is an assumption for Rasch modeling.[36] Rasch analyses were performed using the R package *TAM* (version 3.4),[56] and confirmatory factor analyses were preformed using the R package *lavaan* (version 0.5-12).[57]

Unidimensionality and fit

Using Confirmatory Factor Analysis (CFA), the item data from each individual progress variable were first separately evaluated for unidimensionality and each found to have acceptable data-model fit to a one-factor model. The full set of item data was found to have acceptable data-model fit to a correlated four-factor model, which we believed to be reasonable due to the nested nature of the four progress variables. While discussion of CFA methods for determining the dimensionality of an assessment instrument are beyond the scope of this primer, interested readers are encouraged to review one of the many primary resources for conducting CFAs.[58,59] As this primer focuses only on two items, derived from a single construct map (i.e., Evaluation of Models), we only further discuss those items and their Rasch analysis.

When using Rasch analysis, it is important to evaluate the fit between the data generated from a set of items and the Rasch model, that is whether Rasch is an appropriate tool for interpretation of scores. As noted earlier, it is common to compute both mean square residuals (MNSQ [INFIT and OUTFIT]) and standardized mean square residuals (ZSTD [INFIT and OUTFIT]; the normalized t-score of the residual) as indicators of the extent to which the data fit the Rasch model. Multiple recommendations for “acceptable values” of item fit indices exist in literature; here, we consider MNSQ values between 0.7 and 1.3 and ZTSD values between -2 and 2 to be indicative good model-data fit.[40] The fit statistics for items E1 and E2 (shown in Figure 8) are displayed in Table 1, based on 371 response sets. For both items, all of the fit values, for all steps, are within the acceptable limits. We therefore conclude that Rasch is an appropriate analytic technique and can be used for the interpretation of scores from these items. As Rasch analysis of the full set of metamodeling items will be presented in a subsequent manuscript, only the interpretation of items E1 and E2 (Figure 8) are presented here in this primer.

Table 1. Item fit statistics for items E1 and E2.

Item	Step	Outfit		Infit	
		MSNQ	ZSTD	MSNQ	ZSTD
E1	L0 to L1	1.070	0.053	1.005	0.074
	L1 to L2	0.999	-0.005	0.988	-0.160
	L2 to L3	1.161	1.177	1.035	0.297
E2	L0 to L1	0.953	-0.069	1.001	0.121
	L1 to L2	0.979	-0.187	0.993	-0.040
	L2 to L3	1.005	0.200	1.007	0.271

Horizontal interpretation of items E1 and E2 on Wright map

Ideally, the range of test taker person abilities values and the range of item difficulty values (or in the case of polytomously scored items, step thresholds) overlap to some degree. This overlap provides evidence that the tasks are appropriately difficult for the target population.[40] In the Wright map shown in Figure 10, the person ability values (histogram on left side of Fig. 10) range from a low of around –2.3 logits to a high of +2.7 logits. The logit range of the step thresholds for each answer option (colored symbols on right side of Fig. 10) for item E1 span from around –1.6 to +1.4 logits, with those from E2 spanning from around –2.6 to +0.5 logits. Taken together, these two items are appropriately targeted to this population, but perhaps a bit too easy. This can be seen as the first two levels for each item are below 0 logits. Therefore, only the highest response levels (L3) are difficult enough to discriminate between persons with higher ability values (i.e, those above 0 logits). However, these are only 2 examples of items from the Evaluation of Models scale on the metamodeling instrument. Other, more difficult items are more likely to be useful for discriminating between persons with these higher ability values.

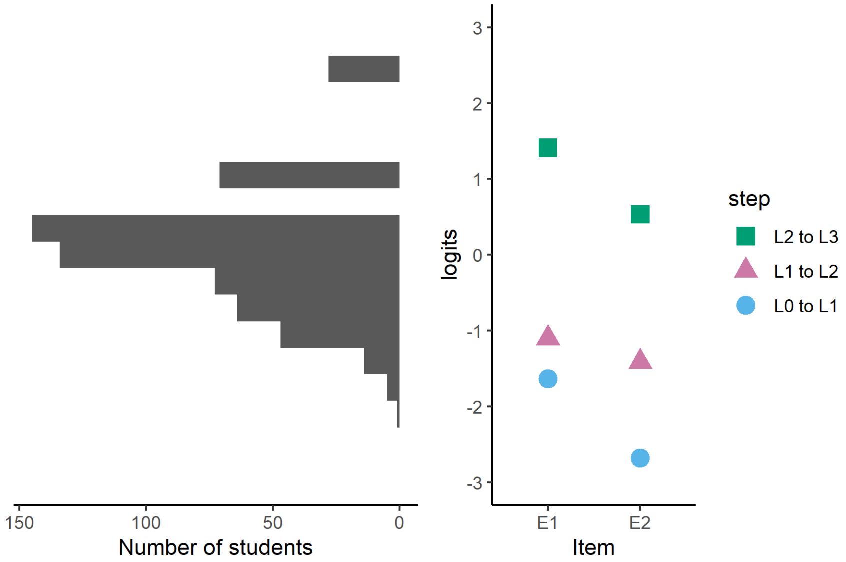


Figure 10. Wright map for items E1 and E2 from the metamodeling OMC assessment. Person ability distribution is shown to the left of the logit scale and item difficulty (Thurstonian thresholds) distribution is shown to the right.

Vertical interpretation of items E1 and E2 on Wright map

In all Rasch analyses, Wright maps provide information regarding the ordering of items by their difficulty level.[40] For OMC items evaluated using the PCM, they are also useful for investigating the 'ordering' of response options within an individual item and therefore relating the response option back to the construct map.[25,31] This part of the analysis is confirmatory in nature, based on the development process and expected ordering of the item response options (informed by the Model of Cognition and Scoring Guide cornerstones) and their corresponding step difficulty parameters. Specifically, it is expected that response options aligned with higher levels on the construct map should have higher response thresholds than response options aligned with lower levels. This can be seen for both item examples in Figure 10, where the step difficulty values from Level 0 to Level 1 (circles) are lower than the step from Level 1 to Level 2 (triangles) and Level 2 to Level 3 (squares). It can also be seen on the Wright map that item E1 is more difficult than item E2, indicated by the higher vertical position of each response option for item E1, relative to item E2.

Item Option Characteristic Curves (IOCCs)

To elaborate on the overlap among person ability values and item response option difficulties, we move away from the Wright map display and turn to Item Option Characteristic Curves (IOCCs; also referred to as category probability curves).[25,38] The IOCC for an item is a plot of the probability [P(θ)] of selecting a given response option as a function of person ability value (θ). IOCCs are useful for evaluating the orderedness of the response options (as are Wright maps) as well as for evaluating whether the response options are useful for grouping students by their ability level values, that is, for distinguishing between qualitatively different groups of students based on the construct map levels. In the IOCC of item E2 (Figure 11), for each response option corresponding to Levels 0-3, there is a range of person ability values (θ) where that response option (i.e., level) is the most probable selection. For instance, a student with an ability level value (θ) of -2 is most likely to select the Level 0 response option, while a student with a slightly higher ability level value (θ) of -1.5 is most likely to select the Level 1 response option and so on.

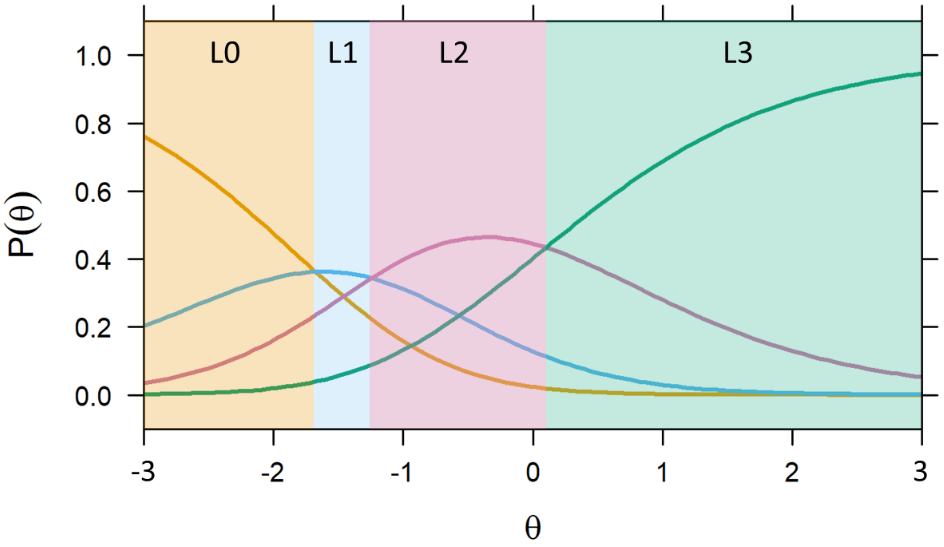


Figure 11. Item option characteristic curve for item E2. Each colored curve represents the probability [P(θ)] of selecting a given response option as a function of person ability (θ) value. The colored boxes correspond to the level of the response option (e.g., L2) which is most probable for a range of θ values.

However, in the IOCC for item E1 (Figure 12), there is no point along the person ability value (θ) range where the Level 1 response option is the most probable selection. While the probability for selecting the Level 1 response reaches its highest value of 0.2 (P(θ)) at an ability value level of -1, students at this ability level value are twice as likely to choose either the Level 0 or the Level 2 response option. Therefore, in its current form and with this population, the Level 1 response option would be considered 'non-functional' and item E1 can only provide information across three of the levels.

The IOCC for item E1 provides a good example for how these curves can be used to also identify items that could be improved through item revision. For instance, at this point in the evaluation of item E1, it would be reasonable to drop the L1 response option and include only three response options. Alternatively, one could investigate, through qualitative methods (e.g., response process interviews), why the L1 response option is not an attractive option for students with low-moderate person ability values (θ) and thereby make and retest an updated version of the response option and item. Additionally, if the L1 response option is shown to be non-functional across multiple items, qualitative investigations could lead to alterations of the construct map itself, thereby feeding back into the evidence supporting the model of cognition.

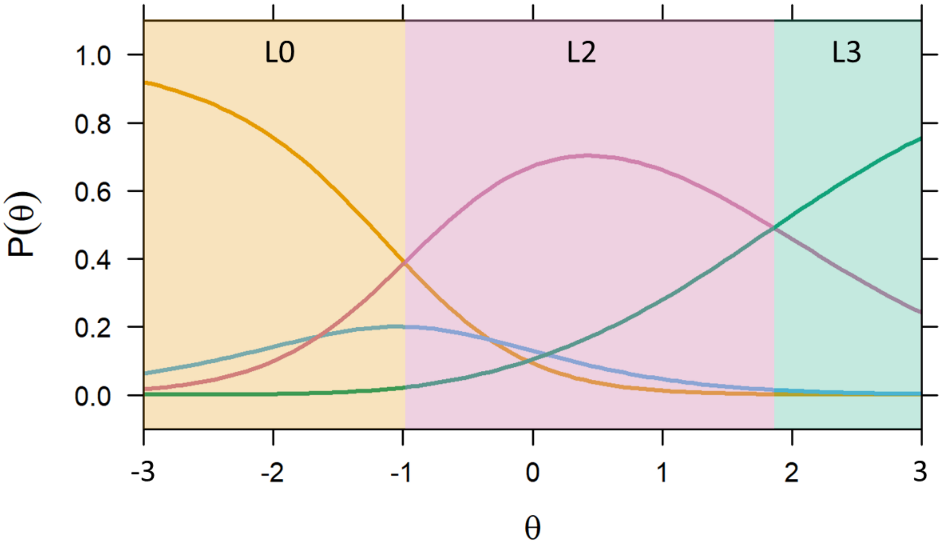


Figure 12. Item option characteristic curve for item E1. Each colored curve represents the probability [P(θ)] of selecting a given response option as a function of person ability (θ) value. The colored boxes correspond to the level of the response option (e.g., L2) which is most probable for a range of θ values.

Summary

Early in this primer we introduced the cornerstone model (Figure 2B) as an assessment cycle that can be used to produce more robust measures.[31,32] A feature of this model that supports the 'robustness' of measures is the interconnectedness among the cornerstones during the assessment development cycle. A good example of this comes in considering how the measurement model (Cornerstone 4) provides feedback for all other cornerstones in the cycle. As stated by Boone et al.[60] in the quote below, Rasch analysis is useful for connecting assessment data to theories of learning. "Rasch measurement is certainly quantitative, but it is also very qualitative... Rasch measurement requires qualitative reflection. Data are not just run through a program and numbers computed to the thousandth place; rather, theory is used to guide an analysis. If theory is not confirmed, then reflection takes place." (Boone et al.,[60] p. 1)

In the context of OMC assessments, analysis of scores using Rasch as the Measurement Model (cornerstone 4) is useful for connecting outcomes back to the other three cornerstones and informing reflection on how students' knowledge, along a developmental continuum, can be assessed. With regard to cornerstone 1, Rasch is used to evaluate whether the orderedness of response options aligns with the theoretical progression developed through qualitative studies and represented by the construct map. With regard to cornerstones 2 and 3, Rasch is used to identify specific items or item response options that do not fit the Rasch model (via fit statistics), which are misordered (via Wright map and IOCCs), or which are not useful for discriminating between students of differing abilities (via IOCCs). These interconnections provide checks across each phase of the development cycle (Figure 2B), leading to more robust evidence to support the outcomes of an assessment and the inferences drawn from them.

QUESTION 3: FOR WHAT PURPOSES ARE OMC ASSESEMENTS USEFUL IN CER?

To support the inclusion of cognitive and measurement theories into assessment instrument development

The development and evaluation of OMC assessments is informed by both a developmental perspective of learning and strong measurement theory. Therefore, these types of assessments can provide support for educators interested in moving away from a deficit model to one of dynamic development and the reweaving of ideas.[3] While a few extant assessments in CER are explicitly linked to developmental perspectives on learning,[18–20,61,62] continued efforts are needed in responding to the call for creation of such assessments.[16] From their inception, OMC items were created as assessment tools capable of linking student cognitive development to measured variables using Rasch modeling.[25] Using the Partial-Credit Model,[38] Rasch allows for the evaluation of multiple-choice items in a polytomous fashion, thereby supporting response interpretation beyond simply correct and incorrect, as employed in classical test theory (CTT).[63] Furthermore, Rasch analysis provides multiple levels of item and test diagnostic information, beyond that of CTT, to provide evidence in support of the outcomes and inferences from an assessment.

OMC assessments provide a formative assessment alternative to the surfeit of concept inventories (CIs) available within CER. Like OMC assessments, CI response options are typically derived based on existing literature or qualitative studies on student thinking but with one of the response options representing the scientifically normative way of thinking. However, OMC assessments are explicitly aligned with developmental process views of conceptual change, which involve the restructuring of knowledge, as opposed to the “misconceptions movement” that has largely informed the development and use of CIs.[64] CIs have been most commonly used to identify misconceptions (or alternative conceptions), which have been described as “*entrenched but false prior beliefs [which] interfere with learning and need to be overcome.*”[64] In contrast, OMC assessments can also be used to identify non-normative conceptions, referred to as “common errors” by Wilson,[7] but these errors in thinking are honored as “*developmental stages, rather than barriers to student progress.*”[25] As such, common errors “*are resolved in the next level of the construct map.*”[25]

While some CIs have been developed and analyzed using more robust measurement models such as Rasch or Item Response Theory (IRT),[65–67] many are supported by CTT analyses only. However, even CIs developed with these more robust measurement models lack strong connections back to a cognitive model of the development of learning. For example, in development and testing of the Thermochemistry Concept Inventory (TCI), Wren and Barbera employed the Partial-Credit Rasch Model and produced a Wright map (like in Figure 10 above) and IOCCs similar to those shown in Figures 11 and 12.[65] However, while these elements help to provide strong evidence for the proper functioning of the items and response options on the TCI, they do not provide an instructor with information about students' level of understanding of the thermochemistry concepts covered by this formative assessment. While CIs have their place in CER studies and classroom-based formative assessment, as tools intended to identify "*the misconceptions that students have about the concepts and principles of chemistry*,"[68] it is common practice to draw inferences based on a single item from a CI.[65,69] In contrast, OMC assessments typically consist of multiple items based on the same construct map, and therefore, inferences about student knowledge are based on this set of items.[18,20]

For evaluating and advancing curricular reform

The linking of OMC items to construct maps results in assessments with the potential to both inform and evaluate curricular reform.[16] Inherent in the construct map approach is the idea that students' learning consists of progression along a progress variable.[7,25] Therefore, this type of assessment data can motivate instructional interventions and curricular reforms with the aim of supporting progression along desired progress variables. In the context of postsecondary chemistry teaching, two curricular reform efforts have focused on narrowing the list of desired progress variables to those essential to a working knowledge of general chemistry. In Chemistry, Life, the Universe, and Everything (CLUE), the progress variables are structure, properties, and energy.[22] Therefore, in the CLUE curriculum, all material centers around these progress variables. In Chemical Thinking, the progress variables are conceived as core concepts that enable chemists to ask and answer questions specific to: chemical identity, structure-property relationships, chemical causality, chemical mechanism, chemical control, and benefits-costs-risks.[70] These examples of curricular changes are especially timely and appropriate for postsecondary chemistry for two reasons. First, introductory chemistry courses serve many students who are not going to be chemistry majors, and second, they are often accused of being a "mile wide and an inch deep."[71–74] For chemistry majors, there is a unique opportunity to build on core ideas throughout the entire degree, which would ideally result in conceptual depth about these core ideas.[48,70,75] Thus, it is critical that the CER community continues to engage in conversations about which progress variables will best support students' use of chemical knowledge in their respective fields.

As the CER community continues to identify progress variables aligned with core ideas in chemistry, it is equally important to have assessments aligned with such learning objectives.[3,16] While Chemical Thinking[23] and CLUE[22] use both formative and summative assessments that are aligned with a developmental perspective on learning, more traditional assessment approaches are commonly used in CER as evidence of the efficacy of reformed curricula.[76–79] For example, the assessments developed by the ACS Exams Institute have been widely used to evaluate reformed curricula.[80–82] However, this approach to curricular evaluation often "*shows only limited effectiveness with traditional, well-established assessments such as standardized tests, because these tests typically address traditional course content and skills.*"[16] Assessments like ACS Exams, which are designed for summative purposes (i.e., making inferences about the overall achievement of an individual) are not always appropriate tools for more formative purposes (i.e., making inferences about learners' existing ideas and the necessary next steps in instruction).[83] Because ACS Exams and other commonly used assessments (e.g., concept inventories and traditional multiple-choice midterm and final exams) are typically scored dichotomously, it may be the case that reformed curricula affect students' knowledge, skills, and abilities but the assessments employed are not sensitive to these changes. Therefore, Holme et al.[16] advocate for using assessments that are more aligned with the goals and purposes of reformed curricula to both evaluate reformed curricula as well as to inform curricular change (Figure 13), though they acknowledge that administration of assessments beyond typical content assessments (e.g., midterm and final exams) require time and resources not always available.

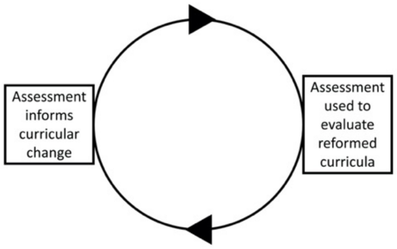

Assessment informs curricular change
Assessment used to evaluate reformed curricula

Figure 13. Depiction of the role of assessment in both informing curricular reform and evaluating reform efforts.

Because OMC assessments are necessarily based on progressions of cognitive skills (e.g., developing particulate-level explanations of chemical phenomena) or concept knowledge (e.g., knowledge about conservation of energy and energy transfer) that develop over time, they are likely to be more aligned with core learning objectives in reformed chemistry courses.[3,16,84] At the K-12 level, the Next Generation Science Standards and partner document, Developing Assessments for the Next Generation Science Standards advocate for assessments based on a developmental perspective that can be used to assess progress toward scientific thinking over time and identify appropriate 'stepping stones' that constitute acceptable and developmentally appropriate intermediate knowledge between naïve conceptions and more advanced conceptions.[85,86]

To provide easy-to-use diagnostic tools for instructors/practitioners

Briggs et al.[25] first introduced OMC format assessments as a potentially "*effective and efficient means of communicating diagnostic results to schools, teachers and students.*" One of the strengths of OMC assessments is that they can be administered and scored rapidly to provide formative diagnostic information for instructors. Information about students' conceptions prior to instruction may be useful for identifying instructional strategies and resources to build on students' existing knowledge. For instance, OMC assessment data can be analyzed at the classroom-level to identify commonly held ideas or beliefs which are scientifically incorrect or naive (common errors) and identify content areas to focus on during instruction in order to resolve common errors. OMC assessment data post-instruction can also be used to provide feedback on students' learning progress, even if students do not reach the highest level (scientifically normative ideas). At the item level, OMC assessment data may also be useful for identifying contexts where students were able to exhibit high-level ideas; instruction can then be designed to leverage students' existing ideas ("conceptual strengths") and help students develop more coherent knowledge across contexts.[25]

In some ways, the diagnostic utility of OMC assessments is similar to that of concept inventories. Some scholars, for instance, Steif and Hansen[87] advocate for the analysis of 'incorrect' answers to CI items to identify specific misconceptions that students exhibit in the context of the CI items; instruction can then be tailored to rectify commonly held misconceptions. However, while in some cases Rasch analysis of distractors on CI items has been useful in identifying answer options that are most attractive to students at different person ability levels (e.g., Wren & Barbera[65]), the "distractor" answer options for CI items are not likely to align across items, and misconceptions are likely to be bound within the context of the specific item. In contrast, data about students' progress along a progress variable is informed by all items within the assessment. Therefore, students' overall person ability level is less context specific.

Well-designed OMC assessments can provide valuable diagnostic information about student knowledge while honoring students' mid-level ideas (between naive and scientifically normative). OMC assessments can be rapidly scored and intuitively interpreted, based on the construct maps from which they are derived. Practitioners can use OMC assessment data to inform instruction and identify contexts which elicit students' "best" ideas and which may be particularly productive for addressing common errors in student thinking.

ADVANCING THE USE OF OMC ASSESSMENTS IN CER

Our intention for this primer on OMC assessments is to provide an additional way facilitate the forward motion being made in the chemistry education community with regard to instructional reforms. As illustrated by the two projects used in this manuscript, there is no single starting point from which a construct map and set of OMC items can be developed. As several lists of core ideas exist in our community,[22,55,70] there are many natural starting points from which to begin building aligned assessments. For example, the Structure and Composition of Matter OMC assessment, developed by Hadenfeldt et al.[18] began with their review of the extensive prior literature in this area. In a similar fashion, one could leverage an extant literature review of a core concept. For example, Bains and Towns[88] published a comprehensive review on the teaching and learning of kinetics, which could therefore be used to provide a similar starting point. In areas such as these, where the research in our community has been fruitful, a new assessment project might begin by the hierarchical arrangement of the extant ideas and insights. An additional starting point that differs from those utilized in our work is to build OMC items around an already developed learning progression (LP). In their Editorial introducing a recent themed issue of *Chemistry Education Research and Practice* focused on Learning Progressions and Teaching Sequences, Bernholt and Sevian note several extant LPs of chemistry concepts.[15] These, as well as other LPs in the special issue would provide an already ordered path from which items and item responses could aligned. No matter the starting point, we encourage other researchers in chemistry education to join us, and those that have preceded us, in continuing to move forward with our assessment practices and aligning them with developmental perspectives of the core ideas in our field.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jack.barbera@pdx.edu

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation's Graduate Research Fellowship (GRFP) Program under award 1650114 and the ECR Core Research: Building Capacity in STEM Education Research (BCSER) Program under award 1937593. 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.

REFERENCES

(1) Taber, K. S. *Modelling Learners and Learning in Science Education: Developing Representations of Concepts, Conceptual Structure and Conceptual Change to Inform Teaching and Research*; Springer Netherlands, 2013.

(2) Taber, K. S. The Nature of Student Conceptions in Science. In *Science Education. New Directions in Mathematics and Science Education*; Taber, K. S., Akpan, B., Eds.; SensePublishers: Rotterdam, 2017; pp 119–131. https://doi.org/10.1007/978-94-6300-749-8_9.

(3) Cooper, M. M.; Stowe, R. L. Chemistry Education Research: From Personal Empiricism to Evidence , Theory , and Informed Practice. *Chem. Rev.* **2018**. https://doi.org/10.1021/acs.chemrev.8b00020.

(4) Driver, R.; Easley, J. Pupils and Paradigms: A Review of Literature Related to Concept Development in Adolescent Science Students. *Stud. Sci. Educ.* **1978**, *5* (1), 61–84. https://doi.org/10.1080/03057267808559857.

(5) Nakhleh, M. B. Why Some Students Don't Learn Chemistry: Chemical Misconceptions. *J. Chem. Educ.* **1992**, *69* (3), 191. https://doi.org/10.1021/ed069p191.

(6) Posner, G. J.; Strike, K. A.; Hewson, P. W.; Gertzog, W. A. Accommodation of a Scientific Conception: Toward a Theory of Conceptual Change. *Sci. Educ.* **1982**, *66* (2), 211–227.

https://doi.org/10.1002/sce.3730660207.

(7) Wilson, M. Measuring Progressions: Assessment Structures Underlying a Learning Progression. *J. Res. Sci. Teach.* **2009**, *46* (6), 716–730. https://doi.org/10.1002/tea.20318.

(8) diSessa, A. A.; Sherin, B. L. What Changes in Conceptual Change? *Int. J. Sci. Educ.* **1998**, *20* (10), 1155–1191. https://doi.org/10.1080/0950069980201002.

(9) Novak, J. D. Meaningful Learning: The Essential Factor for Conceptual Change in Limited or Inappropriate Propositional Hierarchies Leading to Empowerment of Learners. *Sci. Educ.* **2002**, *86* (4), 548–571. https://doi.org/10.1002/sce.10032.

(10) Smith, C. L.; Wiser, M.; Anderson, C. W.; Krajcik, J. FOCUS ARTICLE: Implications of Research on Children's Learning for Standards and Assessment: A Proposed Learning Progression for Matter and the Atomic-Molecular Theory. *Meas. Interdiscip. Res. Perspect.* **2006**, *4* (1–2), 1–98.

(11) Council, N. R. *Systems for State Science Assessment*; Wilson, M. R., Bertenthal, M. W., Eds.; The National Academies Press: Washington, DC, 2006. https://doi.org/10.17226/11312.

(12) Council, N. R. *Taking Science to School: Learning and Teaching Science in Grades K-8*; Duschl, R. A., Schweingruber, H. A., Shouse, A. W., Eds.; The National Academies Press: Washington, DC, 2007. https://doi.org/10.17226/11625.

(13) Kennedy, C.; Brown, N.; Draney, K.; Wilson, M. Using Progress Variables and Embedded Assessment to Improve Teaching and Learning. In *American Education Research Association, Montreal, Canada*; 2005.

(14) Catley, K.; Lehrer, R.; Reiser, B. *Tracing a Prospective Learning Progression for Developing Understanding of Evolution*; 2005.

(15) Bernholt, S.; Sevian, H. Learning Progressions and Teaching Sequences – Old Wine in New Skins? *Chem. Educ. Res. Pract.* **2018**, *19* (4), 989–997. https://doi.org/10.1039/C8RP90009D.

(16) Holme, T.; Bretz, S. L.; Cooper, M.; Lewis, J.; Paek, P.; Pienta, N.; Stacy, A.; Stevens, R.; Towns, M. Enhancing the Role of Assessment in Curriculum Reform in Chemistry. *Chem. Educ. Res. Pract.* **2010**, *11* (2), 92–97. https://doi.org/10.1039/C005352J.

(17) Scalise, K.; Claesgens, J.; Wilson, M.; Stacy, A. ChemQuery: An Assessment System for Mapping Student Progress in Learning Chemistry. In *STEM Assessment Conference*; 2006; p

227.

(18) Hadenfeldt, J. C.; Bernholt, S.; Liu, X.; Neumann, K.; Parchmann, I. Using Ordered Multiple-Choice Items To Assess Students' Understanding of the Structure and Composition of Matter. *J. Chem. Educ.* **2013**, *90* (12), 1602–1608. https://doi.org/10.1021/ed3006192.

(19) Wei, S.; Liu, X.; Wang, Z.; Wang, X. Using Rasch Measurement To Develop a Computer Modeling-Based Instrument To Assess Students' Conceptual Understanding of Matter. *J. Chem. Educ.* **2012**, *89* (3), 335–345. https://doi.org/10.1021/ed100852t.

(20) Park, M.; Liu, X.; Waight, N. Development of the Connected Chemistry as Formative Assessment Pedagogy for High School Chemistry Teaching. *J. Chem. Educ.* **2017**, *94* (3), 273–281. https://doi.org/10.1021/acs.jchemed.6b00299.

(21) Noyes, K.; Cooper, M. M. Investigating Student Understanding of London Dispersion Forces: A Longitudinal Study. *J. Chem. Educ.* **2019**, acs.jchemed.9b00455. https://doi.org/10.1021/acs.jchemed.9b00455.

(22) Cooper, M.; Klymkowsky, M. Chemistry, Life, the Universe, and Everything: A New Approach to General Chemistry, and a Model for Curriculum Reform. *J. Chem. Educ.* **2013**, *90* (9), 1116–1122. https://doi.org/10.1021/ed300456y.

(23) Talanquer, V.; Pollard, J. Let's Teach How We Think Instead of What We Know. *Chem. Educ. Res. Pract.* **2010**, *11* (2), 74–83.

(24) Wilson, M. *Constructing Measures. An Item Response Modeling Approach*; Psychology Press: New York, 2005. https://doi.org/10.4324/9781410611697.

(25) Briggs, D. C.; Alonzo, A. C.; Schwab, C.; Wilson, M. Diagnostic Assessment With Ordered Multiple- Choice Items. *Educ. Assess.* **2006**, *11* (1), 33–63. https://doi.org/10.1207/s15326977ea1101.

(26) Arjoon, J. A.; Xu, X.; Lewis, J. E. Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence. *J. Chem. Educ.* **2013**, *90* (5), 536–545. https://doi.org/10.1021/ed3002013.

(27) Mislevy, R. J.; Almond, R. G.; Lukas, J. F. *A BRIEF INTRODUCTION TO EVIDENCE-CENTERED DESIGN*; John Wiley & Sons, Ltd, 2003; Vol. 2003. https://doi.org/10.1002/j.2333-

8504.2003.tb01908.x.

(28) Mislevy, R. J.; Wu, P.-K. Missing Responses and Irt Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing. *ETS Res. Rep. Ser.* **1996**, *1996* (2), i–36. https://doi.org/10.1002/j.2333-8504.1996.tb01708.x.

(29) Mislevy, R. J. EVIDENCE AND INFERENCE IN EDUCATIONAL ASSESSMENT1,2. *ETS Res. Rep. Ser.* **1995**, *1995* (1), i–44. https://doi.org/10.1002/j.2333-8504.1995.tb01643.x.

(30) National Research Council. *Knowing What Students Know: The Science and Design of Educational Assessment*; Washington DC, 2001. https://doi.org/10.17226/10019.

(31) Brown, N. J. S.; Wilson, M. A Model of Cognition: The Missing Cornerstone of Assessment. *Educ. Psychol. Rev.* **2011**, *23* (2), 221–234. https://doi.org/10.1007/s10648-011-9161-z.

(32) Brown, N. J. S.; Nagashima, S. O.; Fu, A.; Timms, M.; Wilson, M. A Framework for Analyzing Scientific Reasoning in Assessments. *Educ. Assess.* **2010**, *15* (3–4), 142–174. https://doi.org/10.1080/10627197.2010.530562.

(33) Brown, N. J. S. *The Multidimensional Measure of Conceptual Complexity*; Berkeley, 2005.

(34) Harsh, J. A. Designing Performance-Based Measures to Assess the Scientific Thinking Skills of Chemistry Undergraduate Researchers. *Chem. Educ. Res. Pract.* **2016**, *17* (4), 808–817. https://doi.org/10.1039/C6RP00057F.

(35) Wu, M.; Adams, R. J. *Applying the Rasch Model to Psycho-Social Measurement: A Practical Approach*; PANDORA electronic collection; Educational Measurement Solutions, 2007.

(36) Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Danmarks Paedogogiske Institut: Copenhagen, Denmark, 1960.

(37) Downing, S. M.; Haladyna, T. M. *Handbook of Test Development*; Lawrence Erlbaum, 2006.

(38) Masters, G. N. A Rasch Model for Partial Credit Scoring. *Psychometrika* **1982**, *47* (2), 149–174. https://doi.org/10.1007/BF02296272.

(39) Wu, M.; Tam, H.-P.; Jen, T.-H. *Educational Measurement for Applied Researchers: Theory into Practice*; Springer, Singapore, 2016. https://doi.org/10.1007/978-981-10-3302-5.

(40) Bond, T. *Applying the Rasch Model*; 2015. https://doi.org/10.4324/9781315814698.

(41) Stevens, S. Y.; Delgado, C.; Krajcik, J. S. Developing a Hypothetical Multi-Dimensional Learning

Progression for the Nature of Matter. *J. Res. Sci. Teach.* **2010**, *47* (6), 687–715. https://doi.org/10.1002/tea.20324.

(42) Becker, N.; Rasmussen, C.; Sweeney, G.; Wawro, M.; Towns, M.; Cole, R. Reasoning Using Particulate Nature of Matter: An Example of a Sociochemical Norm in a University-Level Physical Chemistry Class. *Chem. Educ. Res. Pract.* **2013**, *14* (1), 81–94.

(43) Adbo, K.; Taber, K. S. Learners' Mental Models of the Particle Nature of Matter: A Study of 16-year-old Swedish Science Students. *Int. J. Sci. Educ.* **2009**, *31* (6), 757–786. https://doi.org/10.1080/09500690701799383.

(44) Cooper, M. M.; Underwood, S. M.; Hilley, C. Z.; Klymkowsky, M. W. Development and Assessment of a Molecular Structure and Properties Learning Progression. *J. Chem. Educ.* **2012**, *89* (11), 1351–1357. https://doi.org/10.1021/ed300083a.

(45) Cooper, M. M.; Underwood, S. M.; Hilley, C. Z. Development and Validation of the Implicit Information from Lewis Structures Instrument (IILSI): Do Students Connect Structures with Properties? *Chem. Educ. Res. Pract.* **2012**, *13* (3), 195–200.

(46) Cooper, M. M.; Corley, L. M.; Underwood, S. M. An Investigation of College Chemistry Students' Understanding of Structure–Property Relationships. *J. Res. Sci. Teach.* **2013**, *50* (6), 699–721. https://doi.org/10.1002/tea.21093.

(47) Becker, N. M.; Cooper, M. M. College Chemistry Students' Understanding of Potential Energy in the Context of Atomic–Molecular Interactions. *J. Res. Sci. Teach.* **2014**, *51* (6), 789–808. https://doi.org/10.1002/tea.21159.

(48) Wolfson, A. J.; Rowland, S. L.; Lawrie, G. A.; Wright, A. H. Student Conceptions about Energy Transformations: Progression from General Chemistry to Biochemistry. *Chem. Educ. Res. Pract.* **2014**, *15* (2), 168–183. https://doi.org/10.1039/C3RP00132F.

(49) R. Zohar, A.; Levy, S. T. Attraction vs. Repulsion – Learning about Forces and Energy in Chemical Bonding with the ELI-Chem Simulation. *Chem. Educ. Res. Pract.* **2019**, *20*, 667–684. https://doi.org/10.1039/c9rp00007k.

(50) Neumann, K.; Viering, T.; Boone, W. J.; Fischer, H. E. Towards a Learning Progression of Energy. *J. Res. Sci. Teach.* **2013**, *50* (2), 162–188. https://doi.org/10.1002/tea.21061.

(51) Lazenby, K.; Stricker, A.; Brandriet, A.; Rupp, C. A.; Mauger-Sonnek, K.; Becker, N. M. Mapping Undergraduate Chemistry Students' Epistemic Ideas about Models and Modeling. *J. Res. Sci. Teach.* **2019**, *57* (5), 794–824. https://doi.org/10.1002/tea.21614.

(52) Gotwals, A. W. Learning Progressions For Multiple Purposes. In *Learning Progressions in Science: Current Challenges and Future Directions*; Alonzo, A. C., Gotwals, A. W., Eds.; SensePublishers: Rotterdam, 2012; pp 461–472. https://doi.org/10.1007/978-94-6091-824-7_19.

(53) Alonzo, A. C.; Steedle, J. T. Developing and Assessing a Force and Motion Learning Progression. *Sci. Educ.* **2009**, *93* (3), 389–421. https://doi.org/10.1002/sce.20303.

(54) Brown, T. L.; Lemay, H.E., J.; Bursten, B. E.; Murphy, C. J.; Woodward, P. . *Chemistry: The Central Science*, 12th ed.; Pearson Prentice Hall: Upper Saddle River, NJ, 2012.

(55) Holme, T.; Murphy, K. The ACS Exams Institute Undergraduate Chemistry Anchoring Concepts Content Map I: General Chemistry. *J. Chem. Educ.* **2012**, *89* (6), 721–723. https://doi.org/10.1021/ed300050q.

(56) Robitzsch, A.; Kiefer, T.; Wu, M. TAM: Test Analysis Modules. R package 2020.

(57) Rosseel, Y. Lavaan: An R Package for Structural Equation Modeling and More. *J. Stat. Softw.* **2012**, *48* (2), 1–36.

(58) Kline, R. B. Principles and Practice of Structural Equation Modeling, 3rd Ed. *Principles and practice of structural equation modeling.* Guilford Press: New York, NY, US 2011.

(59) Brown, T. A. Confirmatory Factor Analysis for Applied Research, 2nd Ed. *Confirmatory factor analysis for applied research.* The Guilford Press: New York, NY, US 2015.

(60) Boone, W. J.; Yale, M. S.; Staver, J. R. *Rasch Analysis in the Human Sciences*; 2014. https://doi.org/10.1007/978-94-007-6857-4.

(61) Scalise, K.; Claesgens, J.; Wilson, M.; Stacy, A. Contrasting the Expectations for Student Understanding of Chemistry with Levels Achieved: A Brief Case-Study of Student Nurses. *Chem. Educ. Res. Pract.* **2006**, *7* (3), 170. https://doi.org/10.1039/b5rp90022k.

(62) Wei, S.; Liu, X.; Jia, Y. Using Rasch Measurement to Validate the Instrument of Students' Understanding of Models in Science (SUMS). *Int. J. Sci. Math. Educ.* **2014**, *12*, 1067–1082.

https://doi.org/10.1007/s10763-013-9459-z.

(63) Crocker, L. M.; Algina, J. *Introduction to Classical and Modern Test Theory*; Cengage Learning: Mason, Ohio, 2008.

(64) diSessa, A. A. A History of Conceptual Change Research. In *The Cambridge Handbook of the Learning Sciences*; Sawyer, R. K., Ed.; Cambridge University Press: Cambridge, 2014; pp 88–108. https://doi.org/10.1017/CBO9781139519526.007.

(65) Wren, D.; Barbera, J. Psychometric Analysis of the Thermochemistry Concept Inventory. *Chem. Educ. Res. Pract.* **2014**. https://doi.org/10.1039/C3RP00170A.

(66) Taskin, V.; Bernholt, S.; Parchmann, I. An Inventory for Measuring Student Teachers' Knowledge of Chemical Representations: Design, Validation, and Psychometric Analysis. *Chem. Educ. Res. Pract.* **2015**, *16* (3), 460–477. https://doi.org/10.1039/C4RP00214H.

(67) Lu, H.; Jiang, Y.; Bi, H. Development of a Measurement Instrument to Assess Students' Proficiency Levels Regarding Galvanic Cells. *Chem. Educ. Res. Pract.* **2020**, *21* (2), 655–667. https://doi.org/10.1039/C9RP00230H.

(68) Bretz, S. L. Designing Assessment Tools To Measure Students' Conceptual Knowledge of Chemistry. In *Tools of Chemistry Education Research*; ACS Symposium Series; American Chemical Society, 2014; Vol. 1166, pp 155-168 SE – 9. https://doi.org/doi:10.1021/bk-2014-1166.ch009.

(69) Mulford, D. R.; Robinson, W. R. An Inventory for Alternate Conceptions among First-Semester General Chemistry Students. *J. Chem. Educ.* **2002**, *79* (6), 739. https://doi.org/10.1021/ed079p739.

(70) Sevian, H.; Talanquer, V. Rethinking Chemistry: A Learning Progression on Chemical Thinking. *Chem. Educ. Res. Pract.* **2014**, *15* (1), 10–23. https://doi.org/10.1039/C3RP00111C.

(71) Cooper, M. Editorial: The Case for Reform of the Undergraduate General Chemistry Curriculum. *J. Chem. Educ.* **2010**, *87* (3), 231–232. https://doi.org/10.1021/ed800096m.

(72) Havighurst, R. J. Reform in the Chemistry Curriculum. *J. Chem. Educ.* **1929**, *6* (6), 1126. https://doi.org/10.1021/ed006p1126.

(73) Lloyd, B. W.; Spencer, J. N. The Forum: New Directions for General Chemistry:

Recommendations of the Task Force on the General Chemistry Curriculum. *J. Chem. Educ.* **1994**, *71* (3), 206. https://doi.org/10.1021/ed071p206.

(74) Nameroff, T. J.; Busch, D. H. Exploring the Molecular Vision: Report from a SOCED Invitational Conference. *J. Chem. Educ.* **2004**, *81* (2), 177. https://doi.org/10.1021/ed081p177.

(75) Fortus, D.; Krajcik, J. Curriculum Coherence and Learning Progressions. In *Second International Handbook of Science Education*; 2012; pp 783–798. https://doi.org/10.1007/978-1-4020-9041-7.

(76) Carmel, J. H.; Jessa, Y.; Yezierski, E. J. Targeting the Development of Content Knowledge and Scientific Reasoning: Reforming College-Level Chemistry for Nonscience Majors. *J. Chem. Educ.* **2015**, *92* (1), 46–51. https://doi.org/10.1021/ed500207t.

(77) Hein, S. M. Positive Impacts Using POGIL in Organic Chemistry. *J. Chem. Educ.* **2012**, *89* (7), 860–864. https://doi.org/10.1021/ed100217v.

(78) McGill, T. L.; Williams, L. C.; Mulford, D. R.; Blakey, S. B.; Harris, R. J.; Kindt, J. T.; Lynn, D. G.; Marsteller, P. A.; McDonald, F. E.; Powell, N. L. Chemistry Unbound: Designing a New Four-Year Undergraduate Curriculum. *J. Chem. Educ.* **2019**, *96* (1), 35–46. https://doi.org/10.1021/acs.jchemed.8b00585.

(79) Perry, M. D.; Wight, R. D. Using an ACS General Chemistry Exam to Compare Traditional and POGIL Instruction. In *Process Oriented Guided Inquiry Learning (POGIL)*; ACS Symposium Series; American Chemical Society, 2008; Vol. 994, pp 20–240. https://doi.org/doi:10.1021/bk-2008-0994.ch020.

(80) Talanquer, V.; Pollard, J. Reforming a Large Foundational Course: Successes and Challenges. *J. Chem. Educ.* **2017**, *94* (12), 1844–1851. https://doi.org/10.1021/acs.jchemed.7b00397.

(81) Lewis, S. E.; Lewis, J. E. Seeking Effectiveness and Equity in a Large College Chemistry Course: An HLM Investigation of Peer-led Guided Inquiry. *J. Res. Sci. Teach.* **2008**, *45* (7), 794–811.

(82) Reid, S. A. Restructuring a General College Chemistry Sequence Using the ACS Anchoring Concepts Content Map. *J. Chem. Educ.* **2020**, 1–8. https://doi.org/10.1021/acs.jchemed.9b00950.

(83) Harlen, W.; James, M. Assessment and Learning: Differences and Relationships between

Formative and Summative Assessment. *Assess. Educ. Princ. Policy Pract.* **1997**, *4* (3), 365–379. https://doi.org/10.1080/0969594970040304.

(84) Stowe, R. L.; Cooper, M. M. Practicing What We Preach: Assessing " Critical Thinking " in Organic Chemistry. *J. Chem. Educ.* **2017**. https://doi.org/10.1021/acs.jchemed.7b00335.

(85) National Research Council. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Idea*; 2012.

(86) Council, N. R. *Developing Assessments for the Next Generation Science Standards*; Pellegrino, J. W., Wilson, M. R., Koenig, J. A., Beatty, A. S., Eds.; The National Academies Press: Washington, DC, 2014. https://doi.org/10.17226/18409.

(87) Steif, P. S.; Hansen, M. A. New Practices for Administering and Analyzing the Results of Concept Inventories. *J. Eng. Educ.* **2007**, *96* (3), 205–212. https://doi.org/https://doi.org/10.1002/j.2168-9830.2007.tb00930.x.

(88) Bain, K.; Towns, M. H. A Review of Research on the Teaching and Learning of Chemical Kinetics. *Chem. Educ. Res. Pr.* **2016**. https://doi.org/10.1039/C5RP00176E.