Journal of Science Education and Technology
https://doi.org/10.1007/s10956-020-09862-4

Combining Machine Learning and Qualitative Methods to Elaborate Students' Ideas About the Generality of their Model-Based Explanations

Joshua M. Rosenberg[1] · Christina Krist[2]

Abstract
Assessing students' participation in science practices presents several challenges, especially when aiming to differentiate meaningful (vs. rote) forms of participation. In this study, we sought to use machine learning (ML) for a novel purpose in science assessment: developing a construct map for students' *consideration of generality*, a key epistemic understanding that undergirds meaningful participation in knowledge-building practices. We report on our efforts to assess the nature of 845 students' ideas about the generality of their model-based explanations through the combination of an embedded written assessment and a novel data analytic approach that combines unsupervised and supervised machine learning methods and human-driven, interpretive coding. We demonstrate how unsupervised machine learning methods, when coupled with qualitative, interpretive coding, were used to revise our construct map for generality in a way that allowed for a more nuanced evaluation that was closely tied to empirical patterns in the data. We also explored the application of the construct map as a framework for coding used as a part of supervised machine learning methods, finding that it demonstrates some viability for use in future analyses. We discuss implications for the assessment of students' meaningful participation in science practices in terms of their considerations of generality, the role of unsupervised methods in science assessment, and combining machine learning and human-driven approach for understanding students' complex involvement in science practices.

Keywords Assessment · Scientific practices · Machine learning · Epistemology · Middle school · Quantitative · Grounded theory · Generality

Given the shifts in science learning and instruction involved, the vision for three-dimensional learning put forth in the Next Generation Science Standards (NGSS; NGSS Lead States 2013) in the United States (U.S.) provide a number of pressing assessment challenges (National Research Council 2014). This vision emphasizes students' *participation in science practices* (Dimension 1) as a means of making progress in building *disciplinary core ideas* (Dimension 3) and using *crosscutting concepts* (Dimension 2; National Research Council 2012) to explain natural phenomena and propose engineering design solutions. Importantly, these three dimensions—practices, core ideas, and crosscutting concepts—are used by students in an integrated way to explain natural phenomena or develop engineering design solutions, an approach known as phenomenon-based teaching (Penuel et al. 2019). This integration requires a movement away from traditional assessment approaches that focus solely on content knowledge.

In this way, the integrated three-dimensional nature (science practices, crosscutting concepts, and disciplinary core ideas) of the NGSS fully embraces the idea—and challenges—of *generality*: Cross-cutting (or generalized) concepts and disciplinary core ideas are to be widely utilized by students in their explanations of phenomena across many different specific content area contexts. At the same time, when students begin with specific contexts—as can be the case with phenomenon-based teaching and learning—it can be

Joshua M. Rosenberg and Christina Krist contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10956-020-09862-4) contains supplementary material, which is available to authorized users.

✉ Joshua M. Rosenberg
jmrosenberg@utk.edu

[1] University of Tennessee, 420 Claxton Complex, 1122 Volunteer Blvd., Knoxville, TN 37996, USA

[2] University of Illinois Urbana-Champaign, Champaign, IL, USA

Published online: 15 September 2020

challenging to generalize from specific cases to broader principles or ideas (Kolodner 1993; Lehrer et al. 2001). Last, practices are meant to be used flexibly and strategically across the varied contexts of scientific activity (Ford 2015; Manz 2015; Lehrer and Schauble 2015).

These goals of generalizability—that students develop cross-cutting knowledge, an understanding of core principles, and the flexible use of practices that can all be leveraged when encountering a new phenomenon or design challenge—suggest that the process of learning science is not only about moving towards a more general understanding but also is about flexibly moving between specific phenomena and more general principles (Berland and Crucet 2016; Tabak and Reiser 1999). In order to do this well, students need to have in mind some recognition of how their science learning lies in regard to its goals for generality (or specificity) as well as how and why they are engaging with those goals. In other words, their engagement in doing science needs to be *meaningful*: they should be meta-aware of the goals for why they are doing what they are doing.

Importantly, *both* the goals of deeply understanding a case and of extracting a general idea are separate from a rote form of participation: following a procedure because that is what the teacher has told us to do today (Berland et al. 2016; Jiménez-Aleixandre et al. 2000). For example, are students trying to deeply understand a specific case because knowing the details helps to articulate how and why the phenomenon occurs? Or, instead, are they identifying the details to fill in the blanks on a worksheet? Similarly, are students trying to understand a specific case so that they can extract from it a more general understanding that will help them explain and predict in more contexts, or are they going through the motions of "discovering" a general rule to recount for an exam? Regardless of whether students are working to generalize the particulars or are applying a general idea, one can imagine both meaningful and rote forms of participation.

Some assessments of three-dimensional science learning aim to assess the dimensions in integration. These efforts have focused on developing construct maps for a single performance expectation (PE: a standard that combines practices, disciplinary content ideas, and crosscutting concepts). While this approach is merited given the complexity of assessing three-dimensional learning, the result of such efforts is rubrics or scoring guides that provide detailed descriptions of what three-dimensional learning might look like in a given content area (e.g., DeBarger et al. 2013; Harris et al. 2019; Penuel et al. 2019), but many do not yet differentiate between meaningful versus rote forms of engagement. Moreover, most focus upon *what* students are doing—albeit with great nuance and value to the field (see Inkinen et al. 2020)—but do not capture the epistemic undergirding of *why* and *how* they are doing it (Manz 2015).

We take the stance that in order to capture meaningful (instead of rote) forms of science learning, we need to measure a construct other than specific content area knowledge or specific practices—even in an integrated way. Instead, we focus on measuring students' *epistemic considerations*: ideas that support and bolster students' participation in knowledge-building practices across content areas. These "how and why" understandings have been described using a range of terms, such as *practical epistemologies* (Sandoval 2005), *epistemic considerations* (Berland et al. 2016), or *epistemic practices* (Kelly 2008; Manz 2015); despite differences in terminology, they are generally accepted as important learning goals for students (Lehrer and Schauble 2015). Notably, with few exceptions (c.f. Sandoval and Millwood 2005), epistemic considerations have been studied through video records (e.g., Krist 2020; Berland and Crucet 2016; Ryu and Sandoval 2012) or domain-general survey instruments (e.g., Kuhn 2000).

In this paper, we present our efforts to assess students' epistemic consideration of generality using embedded assessment questions: those asking students to write about their ongoing classroom activity. Although we have some understanding of how experts consider generality in their scientific practice (e.g., Chinn and Malhotra 2002), we do not yet know what students might be thinking about or considering regarding the generality of their accounts. This is a key assessment issue in that gaining insight into students' thinking about can help us to learn how they might progress towards that expert understanding. Moreover, because most efforts to assess students' epistemic understandings have used video records—which may be necessary for documenting epistemic learning—assessing students' consideration of generality at a scale that can facilitate assessments of many students over a long timescale remains out of reach. This tendency to use detailed video records raises an important assessment-related question: can we assess epistemic considerations in students' written scientific work? Thus, our goal is to elaborate—to work out in detail—the construct of students' consideration of generality, which we argue could provide insights into how to better measure students' meaningful participation in science practices.

To do this, we leverage machine learning (or ML) methods as a tool for inductive pattern seeking of students' consideration of generality while they are constructing model-based, scientific explanations. Specifically, our goal is to strategically use ML methods in conjunction with human coding to elaborate on the ideas students draw upon when we ask them to consider the generality of their model-based explanations. Our approach differs from many current applications of ML in science education that use ML to increase the efficiency or accuracy of coding once a scheme is already well defined (e.g., Gobert et al. 2015; Gerard, and Linn 2016; Nehm et al. 2012; Pei et al. 2019). By focusing on the use of ML for the

elaboration of a construct, we aim to advance a methodology that can be used at a different, earlier point in the analytic process. Thus, our research question is as follows: How can an approach that integrates ML methods and interpretive qualitative coding be used to elaborate students' consideration of generality as a means of assessing students' participation in science practices?

Literature Review

Assessing Participation in Constructing Model-Based Explanations of Phenomena

We focus on the science practice of *constructing model-based explanations* as central to science education (Schwarz et al. 2017). Although *Developing and Using Models* and *Constructing Evidence-Based Explanations* are delineated as distinct practices in the NGSS (NGSS Lead States 2013) and the *Framework for K-12 Science Education* (National Research Council 2012), they are often intertwined or used iteratively since both practices share a goal of explaining the natural world (Lehrer and Schauble 2006; Passmore et al. 2017). In this sense, we are interested in the *epistemic dimensions* undergirding students' participation in knowledge-building—how and why they are justifying their claims, shaping the nature of their account, or determining a level of generality or specificity (Berland et al. 2016)—regardless of the particular structure or form of the knowledge product (i.e., the model or explanation).

The most widely cited learning progression (LP) for modeling was developed by Schwarz et al. (2009); it is one that also is explicitly epistemic in nature. Rather than focusing on the structural forms of students' models, their progression captures whether and how students understand models as *generative tools for predicting and explaining* and *as changeable entities*. These two dimensions make how students are considering various epistemic criteria as they make progress in each dimension explicit. For example, as students make progress in understanding models as changeable entities, the LP articulates differences in how they are considering the *explanatory nature* of their model in terms of whether they are making modifications with attention to how the explanatory power is improved. The LP also articulates differences in how students are *justifying* their models and, importantly, differences in how students are considering the *generality* of their model: whether it is a literal illustration of a specific phenomenon or whether it explains answers to questions about a group of phenomena.

Other existing LPs for specific practices tend to focus more on the structural forms of the products. Also in terms of modeling, Zangori et al. (2015) present an LP that identifies five structural components that students' models should ideally include the following: components, sequence, explanatory process, mapping, and principles. More sophisticated models include more of these components. Gotwals and Songer (2013) elaborate an LP for constructing evidence-based explanations using the popular *Claim-Evidence-Reasoning* (CER) instructional scaffold (McNeill et al. 2006). The levels of this progression capture the extent to which students produce various elements of the explanation (claim, evidence, and/or reasoning) independently or with varying degrees of support.

However, if we look closely, the epistemic dimensions driving the utility of these structural forms are implicit in these LPs. The Gotwals and Songer (2013) progression attempts to capture how students consider *justifying their accounts* (i.e., considering evidence) as part of their explanation construction, and the Zangori et al. (2015) progression captures how students consider *generality* based upon whether students include a scientific principle that links the other parts of their model together, as well as how students consider the *mechanistic nature* of their model by the extent to which they include an articulation of an explanatory process.

Thus, across LPs for the practices of developing models and constructing explanations, we see attention to how students are considering the epistemic criteria for justification, generality, and the mechanistic nature of their accounts, even when these epistemic considerations are not formalized within the LPs. Importantly, these progressions also implicitly assume that what is shifting is *how students are considering* epistemic criteria in their development and use of (model-based) explanatory accounts. An emphasis on a shift in how students are considering various epistemic criteria suggests that analytic attention to the *form or structure* of the knowledge product is insufficient for capturing students' understanding of the epistemic utility or rationale for developing that knowledge product. For example, it is possible for students to appropriate the *forms* of models or explanations (e.g., "I included evidence") without necessarily understanding the epistemic utility for developing these knowledge products (e.g., "Because the worksheet had a slot for evidence." (Berland et al. 2016; Gotwals and Songer 2013). Thus, simply looking at the product itself may not provide the grounds for researchers to understand students' participation in science practices.

Using Epistemic Criteria as a Metric for Students' Participation in Science Practices

We propose using students' consideration of epistemic criteria as the focus for assessment of students' participation in science practices. There is increasing evidence that students' consideration of various epistemic criteria can shift over time through sustained participation in science learning. For example, studies of science classrooms have shown how students

develop epistemic ideas about *justification* and evidence both within a single content-area unit (Manz 2012; Ryu and Sandoval 2012) and across multiple science units (Krist 2020). Similarly, empirical studies have characterized students' epistemic ideas about the *nature of the accounts* they construct in terms of their mechanistic reasoning. The mechanistic sophistication of students' accounts tends to increase over the course of a single content-area unit (e.g., Dickes et al. 2016; Duncan and Tseng 2011). In addition, there is some evidence that these increases in mechanistic reasoning persist over time, even across different content area units (Authors, Krist 2020; Reiser et al. 2016), suggesting that students are learning something beyond specific content knowledge that supports mechanistic account construction.

What is less established through past research—although still implicit in the LPs for science practices described above—is how students develop ideas about the *generality* of the knowledge products they consider. In other words, when we ask students to consider how general (or specific) their models or explanations are, what do they consider? While the philosophy and science studies literatures have articulated how generality is used as a criterion in professional science (Giere 1988; Popper 1959; Thagard 1978), and science education scholars have articulated an "upper anchor" for what considering generality might look like in classrooms (e.g., Berland et al. 2016; Chinn and Malhotra 2002), we have not yet documented how students consider generality, or what it might look like for them to be in the process of shifting towards this more sophisticated understanding. This "loose description" of competencies or outcomes is insufficient for robust assessment efforts (National Research Council 2012).

Given this insufficient understanding of students' consideration of the generality of their scientific work, we situate our work as early in the process of elaborating a *construct*: a focused set of knowledge, understanding, and capabilities that an assessment is designed to measure that can then be used in later assessment development efforts (National Research Council 2014). We draw on what could be described as early stages of *construct-centered design* (Shin et al. 2010) or *construct modeling* (Wilson 2004) approaches, both of which begin with the development of a construct map by defining and unpacking the construct of interest in order to articulate a working definition of what is to be assessed (Morell et al. 2017; National Research Council 2014; Zangori et al. 2015). In defining and unpacking generality, we are taking a "bottom-up" approach that explicitly attends to—and attempts to build upon—learners' initial conceptions (Morell et al. 2017) in a way that could contribute to the subsequent development of an LP for generality.

The Potential of Machine Learning Methods for Inductively Elaborating a Construct

ML techniques can be particularly useful at an early stage of the assessment development process, such as at the stage of elaborating a construct. Machine learning (ML) methods are often considered as supervised (utilizing data and models with an already known outcome or label) or unsupervised (utilizing data and models for outcomes or labels which are not known; Hastie et al. 2009). In science education, ML has been used for a variety of science education assessment-related purposes, such as to increase the efficiency and accuracy of the coding of qualitative (e.g., written or image based) assessment item responses (Gobert et al. 2015; Gerard, and Linn 2016; Nehm et al. 2012; Pei et.al 2019; Zhai et al. 2020). For instance, Gobert and colleagues used log-trace data to identify dimensions of students' planning and carrying out investigations (Gobert et al. 2015; Gobert et al. 2013). Similarly, scholars have examined how supervised ML methods can be reliably used to automatically code students' written explanations and diagrams and models for scientific phenomena (e.g., Gerard and Linn 2016).

ML has also been used in science education as a means to enhance construct or content validity (e.g., Anderson et al. 2020; Beggrow et al. 2014; Sherin 2013). For instance, Sherin (2013) advocated the role of using "human-based and computational methods in tandem, in a manner that increases our confidence in both" (p. 602). He found that unsupervised ML methods (married with Natural Language Processing [NLP] techniques) effectively reproduced a human-based analysis of themes among students' explanations. In another example, Beggrow et al. (2014) showed how an ML approach (combined with NLP) coded students' ideas about scientific phenomena in a way that aligned with the results of human-driven, qualitative coding.

Although these two examples reflect different goals, they use similar input, namely the *content* of answers or explanations from students. None of these efforts have focused on students' espoused *epistemic understandings*. This issue is compounded by the absence of explicit rubrics (described in the last section) for students' epistemic considerations from past research.

In this study, we sought to leverage the unique affordances of ML methods to assess students' epistemic ideas about generality as an important goal for science learning. Specifically, we capitalize on ML's strengths in (a) efficiently analyzing large, qualitative data sources, and (b) revealing new dimensions of the construct being studied. In particular, we focused on the use of unsupervised ML methods. Although they are less used in science education than supervised methods (Zhai et al. 2020), we conjectured they would be useful for illuminating levels or gradations students' ideas about generality

without requiring a precise a priori ordering or leveling of the ideas.

Thus, we were interested in exploring the potential of ML at an earlier stage in the data analysis process, given the potential of ML—especially unsupervised ML—for exploring and identifying underlying patterns in data. Consequently, we adopted a *computational grounded theory* (CGT) approach which coordinates human-driven, qualitative grounded theory analysis with ML methods based on the unique strengths of each (Nelson 2020). This approach aims to simultaneously make the methodological decisions guiding grounded theory more transparent (and reproducible) and to provide more stringent guidelines for qualitatively interpreting meaningful and valid patterns in data using ML methods. The goals of CGT are distinct from other uses of ML early in the analytic process (e.g., Beggrow et al. 2014; Sherin 2013) in that CGT is associated with an expectation that combining human and ML methods in strategic ways can produce different results than would likely be produced by using either alone. Accordingly, we adopted a CGT approach with the aim of grounded conceptual development: to elaborate and iteratively revise a construct map that represented how we characterized students' consideration of the epistemic criterion of generality when constructing model-based scientific explanations. This paper presents a case of using NLP-based ML as part of a CGT approach to elaborate a construct map and to explore that construct map's viability as a framework for coding.

Method

Participants

The participants in this study were 845 middle school students (participating longitudinally from the 6th through 8th grades) from six schools in two different states in the Midwest U.S. The students in all participating schools were using the *Investigating and Questioning our World through Science and Technology* (IQWST) curriculum (Krajcik et al. 2011), a comprehensive science curriculum designed to support students' meaningful participation in science practices (Krajcik et al. 2008). This study was a part of a larger project, Supporting Scientific Practices in Elementary and Middle School Classrooms, which aimed to characterize students' involvement in science practices in classroom contexts.

Curriculum and Context

The IQWST curriculum consists of twelve content-specific units, three each in Physical Sciences, Introduction to Chemistry, Life Sciences, and Earth Sciences. The units are organized around a driving question, such as "How can I smell things from a distance?" In each unit, students investigate phenomena and gradually build and refine models and explanations of those phenomena, culminating in a set of general ideas (or principles) that answer the driving question.

Teachers in the study had varying degrees of familiarity with the IQWST curriculum. Each teacher participated in at least one professional development session provided by members of the research team. Table S1 displays information about each school's enactment of each unit.

Data Sources

To collect information about students' epistemic consideration of generality, we administered embedded assessments to the same cohort of students at multiple time points over three subsequent years. The assessments were embedded in that they asked about models or explanations that students created as a part of typical classroom work. We selected activities from the curriculum materials in which students were synthesizing ideas from several investigations in the form of a model or an explanation. We then embedded questions about these models or explanations into their classroom activities. One of these questions was explicitly designed to elicit students' rationales related to the generality of their model or explanation (see Fig. 1).

For the analyses presented in this paper, we selected four embedded assessment items from two units in the same science content area (6th- and 7th-grade Chemistry) and one in a different content area (8th-grade Life Sciences). We chose these units in order to gradually expand the variation of responses for the NLP-based, ML methods that we used given the potential importance of specific terms which may appear in some units but not others.

In total, the 845 participants completed 1885 embedded assessment responses from across all five assessment administrations. 29.2% ($n = 247$) responded to one item; 35.9% ($n = 303$) to two; 21.8% ($n = 184$) to three; 8.9% ($n = 75$) to four; and 4.3% ($n = 36$) to all five. Missing responses were the result of absences, incomplete assessments, or the unit not being enacted, as not all teachers completed all units (see Table A1).

Data Analysis

The data analysis involved the use of unsupervised and supervised ML methods, as well as human-driven interpretive analysis. We combined these analytic methods by moving from (1) unsupervised ML to (2) human-driven qualitative analysis to (3) supervised ML.

Discovering Student Ideas about the Generality of their Written Model-Based Scientific Explanations Using Unsupervised ML Because a key affordance of unsupervised ML is the ability to detect patterns that may differ from the

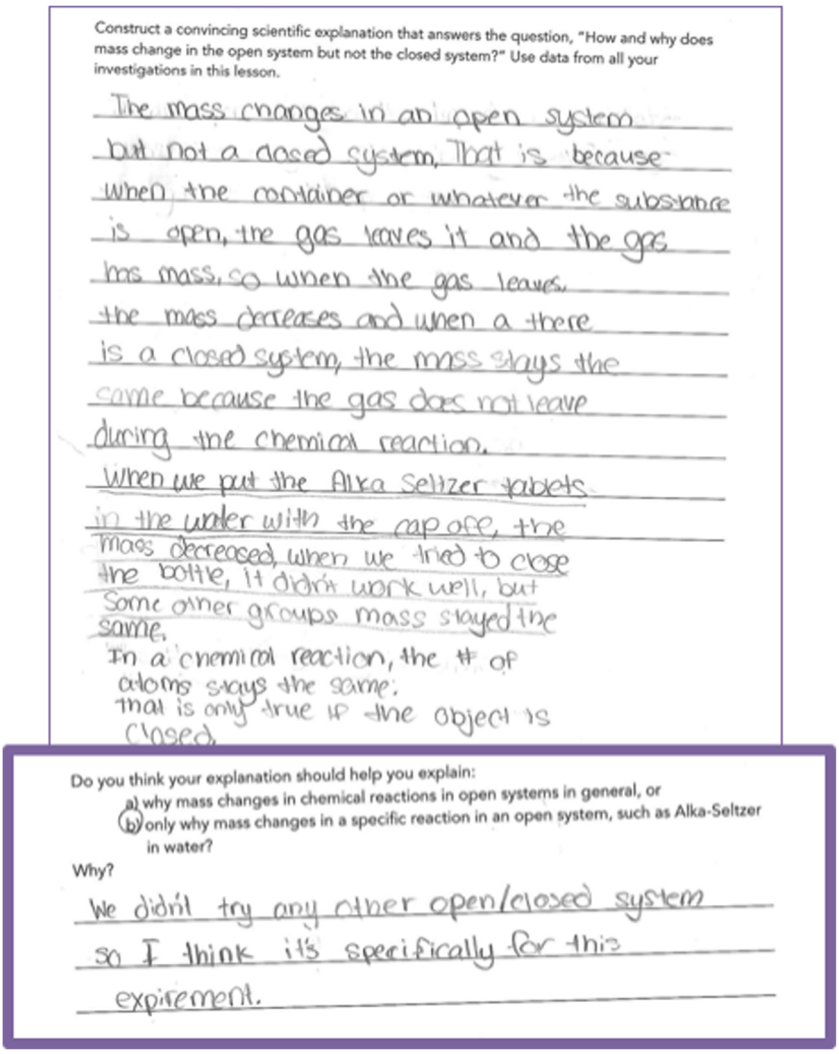

Fig. 1 Explanation prompt and embedded assessment question targeting generality .Note. The purpose box contains the embedded assessment question targeting students' epistemic consideration of generality

ideas that humans expected and therefore tend to find, we first generated a theorized construct map for generality (Table 1). We used this construct map as initial data about what students' written responses were about as well as to theoretically ground pattern interpretation from unsupervised ML analyses.

For this unsupervised ML analysis, we used a subset of 173 student written responses from the 7th-grade Chemistry assessment items. These responses were manually transcribed and stored in a text file. We processed the documents by tokenization with unigrams, removal of stopwords using a common English language dictionary (Benoit et al. 2019b), and stemming using the Snowball C stemmer (Bouchet-Valat 2014). We then created a document-term matrix, a data structure commonly used for natural text data (Hirschberg and Manning 2015). This document-term matrix had 173 rows (documents) and 345 columns (features).

We then used a combined hierarchical agglomerative and *k* means clustering technique (Bergman and El-Khouri 1999)[1] using the log of the term frequencies.[2] To determine the optimal number of clusters, we balanced concerns of interpretability and parsimony with measures of fit following a process similar to that described by Sherin (2013), who aimed to balance such tradeoffs (between interpretability and parsimony), settling on a seven-cluster solution as one which seemed to "resolve interesting features of the

[1] The approach we used has been shown to lend greater stability to the *k* means clustering solution, which can be influenced by the starting points for the algorithm. This approach uses the results from hierarchical clustering as the starting points for *k* means (Bergman and El-Khouri 1999). In our technique, what is being clustered is the vector space representation of each document: in other words, the raw data for the clustering procedure is a row in a table, with values ranging from zero to the maximum number of times any term appears across all documents. The default distance metric for the hierarchical clustering is cosine similarity.

[2] The R package we created and used (Rosenberg and Lishinski 2018) is available to anyone via GitHub for anyone seeking to carry out a similar two-step cluster analysis in R (R Core Team 2019); Sherin (2020) provides a very similar package in python.

Table 1 Theoretical construct map for students' epistemic consideration of generality

Code category	Code	Code description
A. Single level (either general or specific)	A1. No rationale included	Response indicates that the explanation is either general or specific.
	A2. Rationale included	Response indicates that the explanation is either general or specific with a rationale for why one is suitable or better than the other.
B. Level-crossing	B1. Generalizing from specific case to a class of ideas	Response describes how the explanation about a specific case (or phenomena) can apply to a general scientific idea or principle.
	B2. Applying a general idea to a specific case	Response describes how a general scientific idea or principle can be used to explain a specific case (or phenomena).
	B3. Boundary conditions of generality or specificity included	Response describes the conditions under which the explanation does (or does not) apply.

data while producing results . . . that are not overly difficult to interpret" (p. 621).

To interpret the clusters, we used both the most frequent terms and the responses associated with each cluster in order to understand what ideas about generality were evidenced by students' responses. We found that a nine-cluster solution explained the data best.

Interrogating Patterns in Students' Ideas Using Interpretive Qualitative Methods to Create a Construct Map A major benefit of using CGT is that it allows for the identification of unexpected or surprising patterns that may have been overlooked by human coders due to expectation bias or interpretive fatigue (Nelson 2020). The nine clusters that we used did illuminate surprising patterns that differed from our original expectations (see Table 1 above). Accordingly, using a constant comparative qualitative coding approach (e.g., Fram 2013), we interrogated these surprising patterns.

Specifically, we conducted the first round of descriptive coding followed by iterative code mapping (Saldaña 2016) to identify new, theoretically informed categories based on new ways that the responses had been assigned to clusters. As part of this coding, we compared the cluster assignment for each response to how we would have categorized the response based on the theorized construct map for generality. In this process, we annotated and further characterized each response, noting clusters containing group responses that fit a category from the theorized construct map; clusters that contained two or more groups of responses when aligned with the theorized construct map categories; and instances in which groups of responses that fit a theorized construct map category were distributed across multiple clusters. We then determined which of these groupings was meaningful in terms of telling us something about how students were thinking about generality, rather than capturing some other feature of the group of responses. We then conducted an additional round of pattern coding (Saldaña 2016) to develop concise descriptions of the observed patterns. Through these multiple rounds of coding and discussion, we developed a construct map that reflected these meaningful categories.

Exploring the Viability of the Construct Map as a Framework for Coding Last, we explored the viability of the construct map as a framework for coding. Our purpose was not to demonstrate that the construct map could be used for automated coding at scale, but rather to provide some initial validity evidence for the refined coding categories (Nelson 2020; Sherin 2013). We first double coded all 173 responses with the new construct map categories, discussing any disagreements (and clarifying the definitions of the construct map's categories) until 100% agreement was reached. We then used three different algorithms to determine how reliably the codes that resulted from the application of the construct map could be utilized as a part of a supervised ML technique. We used the *quanteda.classifiers* R package (Benoit et al. 2019a) for implementations of each of the three classification algorithms, from the relatively simple (but usually performant for text data) Naïve Bayes (native to the quanteda package), to fairly sophisticated: a sequential neural network (via the keras deep learning library; Allaire and Chollet 2019), along with a support vector machine (from the LiblinearR package; Helleputte 2017). To determine the accuracy of the classifier, we used a Leave One Out Cross Validation (LOOCV) procedure, a procedure which involves using all of the observations in a dataset except for one that is "left out" to train a classifier (or any statistical or ML model) to predict the code of the left out observation.[3] procedure is repeated for every observation in the dataset, such that every observation is predicted using all of the other observations for the purpose of training the classifier. Finally, the agreement between the predictions for every observation obtained through this process and the known values for each observation is calculated. To determine the agreement, we calculated the percentage agreement and quadratic weighted kappa (Cohen 1968) values for the predicted versus actual codes from the LOOCV

[3] LOOCV is equivalent to k folds cross-validation when k is equal to the number of observations in the dataset.

procedure on the initial set of 173 responses. Finally, to explore how accurately the coding frame could be utilized at a larger scale, we coded the remaining responses from the full set of 1885. In this way, we used the smaller ($n = 173$) set of responses to gain an initial understanding of how accurately the coding frame could be used as a proof of concept, and the larger set in order to begin to understand how effectively the construct map could be used at scale (and across multiple content areas).

Findings

Exploring Students' Ideas about Generality

To interpret the nine clusters generated by the unsupervised ML analysis, we examined the most common words in each cluster, read the responses assigned to each cluster, and qualitatively developed an overall theme for each cluster (Table 2). After removing two ambiguous clusters (2 and 9) and collapsing two clusters that were thematically similar (7 and 8), we used the remaining six clusters to refine our construct map.

Interpreting the Clusters and Developing a Construct Map for Use as a Framework for Coding

Next, we leveraged qualitative analysis tools to interpret the six clusters in light of our theorized construct map, with the aim of developing a construct map that reflected both theoretical aspects of generality and the themes in students' responses represented by the results of the cluster analysis—and could potentially be used as a framework for coding. Specifically, we conducted the first round of descriptive coding followed

Table 2 The themes, most common terms, and representative responses for the nine-cluster solution

Cluster Theme (Rationale is about . . .)	Most Common Words	Representative Responses
1: Helping the reader to understand why general or specific would be better	substances, interact, different, need, form, with, acetic, other, stuff, have, those, each, ones, important, any	"Because different substances may cause different results" "Because not all substances form something new, and not all substances have the same reactions in chemical reactions, so it can't explain a general way substances interact and form new ones. Since this project focuses on copper and acetic acid, the model should explain and focus on specific substances, the copper and acid."
2: Unclear/nonsensical	yes, good, has, millions, situations, now, makes, thats, as, wont, general, accurate, each, different, specific	"Yes because" "Yes because general is now specific"
3: The clarity and utility of the representation when it is either general or specific	specific, better, atoms, understand, helps, focuses, describe, us, certain, represent, used, molecules, crowd, easier, want	"It is a new substance because the atoms are the same, just rearranged." "We used the specific molecules and atoms to represent a certain model."
4: Showing how and why the focus of the model form and change	hows, substance, green, my, form, penny, trying, formed, they, are, find, out, b, acetate, react	"The model shows not only how it works for these molecules but how it will work for other molecules." "Because you're trying to know how and was it that made the penny green."
5: Weighing generality or specificity against each other	then, if, just, your, focus, general, was, our, things, than, explains, any, situation, interacts, exactly	"Because if we didn't, then we wouldn't know about what was happening." "We should focus on this because if we draw general models we won't know what we are mixing and what the product is."
6: Similarities and comparisons across processes, mass, theories, and reactions	reactions, mass, other, sense, system, systems, process, same, theories, changes, did, and, does, do, general, evidence	"The process is the same for all chemical reactions, so it wouldn't make sense to model each chemical reaction." "I think this [it should be general] because first of all, scientific principles are demonstrated in different situations, also my claim only says 'In an open system but not in a closed system' and mentions nothing about alka-seltzer."
7: Communicating the "main point" of the task or lesson	show, its, thats, people, everything, could, point, answer, b, experiment, thing, describes, idea, models	"Because that's the whole point of the model." "To show what is going on in the experiment."
8: Interpreted with Cluster 7	were, happened, will, focusing, studying, with, air, are, than, touching, vinegar, without, no, substances, wanted, rain	"I think that because that is what we were studying." "Because we're focusing on more than one chemical."
9: Unclear/nonsensical	why, lower, cemig, rauted, complicating, gets, or, stay, ca, project, since, specific, data, explanation, phonomenon	"Because the cemig rauted."

n denotes the number of responses associated with each cluster

by iterative code mapping (Saldaña 2016) to identify new, theoretically informed categories based on new ways that the responses had been assigned to clusters. We then conducted an additional round of pattern coding (Saldaña 2016) to develop concise descriptions of the observed patterns. The resulting construct map is displayed in Table 3.

As we engaged in these qualitative coding activities, we made several categorizations that differed from our initial construct map. First, we distinguished between responses that had been characterized as a "single level" (category A in Table 1) in a new way. Instead of simply indicating *whether* a response included a rationale, for instance, we coded for specific possible single-level rationales, including communicating clearly or articulating a mechanism. Second, we found that the codes in the theorized construct map for "level-crossing" (category B in Table 1) were not distinguished by the direction of the reasoning students were exhibiting (i.e., whether they were generalizing a specific idea or applying a general idea to a particular case). Instead, students tended to see the task as a general/generalizable one despite the specific nature of its context (code 4A in Table 3); defend their choice of either generality or specificity by negating the other option (code 4B in Table 3); state *that* their response could be used to explain or predict in other situations (code 4C in Table 3); or articulate *how* their response could be used to explain other situations by showing how the mechanism was generalizable (code 5 in Table 3). Last, a group of responses evidencing mechanistic (or "how and why") thinking was not a part of our theorized construct map but was identifiable in one of the clusters (code 3 in Table 3).

The Viability of the Initial Construct Map as a Framework for Coding

After developing the construct map, we explored its viability as a framework for coding the subset of 173 responses using supervised ML methods. We found that the LOOCV-weighted kappa ranged from .47 (for the Naïve Bayes algorithm) to .56 (for the support vector machine; Table 4), indicating *moderate agreement* with the manual codes (Landis and Koch 1977). We then calculated the LOOCV percentage agreement and weighted kappa for this larger set of human-coded responses ($n = 1885$) finding that it ranged from .62 (Naïve Bayes) to .66 (support vector machine; Table 4), indicating substantial agreement with the manual codes (Landis and Koch 1977). Thus, the accuracy increased with the use of additional coded data and was found to be best for the support vector machine classification algorithm. This result is promising for the potential of conducting automated coding using the construct map as a framework for coding, although it will likely require more sophisticated techniques for algorithm refinement.

Discussion

We used ML early in the assessment development process to develop and revise a construct map for students' consideration of generality. We did so by strategically combining unsupervised ML methods with interpretive, human-driven coding. The methodological approach used in this paper aligns with

Table 3 Construct map for epistemic consideration of generality

Construct level	Construct level subtheme "My response being specific or general is better because. . ."	Construct level description
0. Not codable	*Not codable*	Response is blank, unclear, or incomplete.
1. Literal Task Goal	"That's what we did"	Response emphasizes the alignment of explanation and the worksheet instructions or activity task.
2. Communication	A. "It is more 'right'"	Response emphasizes criteria such as clarity, detail, or accuracy of information.
	B. "It supports learning or understanding"	Response emphasizes an audience's knowing, thinking, or the need to be convinced.
3. Mechanism	"It represents the mechanism"	Response emphasizes that it is better to show or explain a phenomenon in terms of how and why something occurs.
4. Generality	A. "The goal was to understand something general about this idea"	Response identifies the bigger picture learning goal (beyond the workbook's instructions)
	B. "It should be A, because it should not be B" (where A is either general or specific, and B is the opposite)	Response is part of an argument making an implicit claim against either generality or specificity.
	C. "It applies, generates, or predicts (or not)"	Response involves an argument for the generality or specificity of their product because it can help them explain or predict more situations.
5. Generality + Mechanism	"The mechanism is generalizable"	Response identifies that it is the process or mechanism that is general; demonstrating how that mechanism transfers.

Table 4 The reliability of the application of supervised ML classification algorithms

n	Classification algorithm	Percentage agreement	Quadratic Weighted Kappa
169	Naïve Bayes	.60	.47
	Support vector machine	.59	.56
	Sequential neural network	.56	.53
1885	Naïve Bayes	.66	.62
	Support vector machine	.70	.66
	Sequential neural network	.70	.65

These values compare the predictions from each of the supervised ML classification algorithms to the codes from human-based coding

broader calls for integrating ML into educational research methods in a way that is responsive to the aims of educators (e.g., Shaffer 2017) and the push for values-driven ML applications (e.g., Greene et al. 2019). In this section, we discuss our key findings in light of research on efforts to integrate ML into science education assessment.

Computational Grounded Theory and Using Unsupervised Machine Learning Methods

While ML is becoming more common for educational assessment and measurement (Burrows et al. 2015), applications of unsupervised methods are, at present, relatively uncommon (Zhai et al. 2020). One reason for this may be the absence of guidance about how they can be used to reach assessment-related ends. We presented a proof-of-concept example of a way to integrate unsupervised ML methods in a theoretically grounded way. Thus, unsupervised methods may be especially valuable as a part of a larger ML process, one which involves not only unsupervised and supervised methods (which but a few science education studies have done; e.g., Wiley et al. 2017), and human-based, qualitative methods. Such a process can lead to codes that reflect both empirical patterns in the data and the influence of the theory, prior knowledge, and value-laden priorities and insights that human analysts bring to interpreting data.

The development of a construct map is an important and challenging part of the broader assessment process (DeBarger et al. 2013; Harris et al. 2019). To address this challenge, we used an unsupervised ML method, cluster analysis (Hastie et al. 2009), to identify clusters of responses. While some of these clusters reflected our original theoretical frame, others reflected new ways of parsing the theoretical categories or new categories altogether. These new patterns allowed us to elaborate and to better "flesh out" the coding scheme to say more about the range of ways that students considered the generality of their explanations.

A few existing studies have used similar approaches. Sherin (2013) carried out a similar cluster analytic approach as a means of providing evidence for the validity of the original qualitative coding of students' conceptual understanding of seasons. Similarly, Anderson et al. (advance online publication) used an NLP-based, topic modeling approach to establish the content-related validity of science education assessment items, and Zehner et al. (2016) compared human and machine coding of open-ended written responses from the *Programme for International Student Assessment (PISA*) assessment. These studies all use computational methods to provide validity evidence for qualitative coding.

What distinguishes our use of ML is our focus on *developing* a construct map. In addition, we were looking at students' epistemic, rather than conceptual, ideas. To our knowledge, this is the first examination of whether NLP methods can be used to identify epistemic themes in students' responses, which are abstract and not necessarily particular to content- or subject- matter vocabulary. We demonstrated that they can, with moderate (with a subset of the data) and substantial (for the larger set of coded data) accuracy.

We posit that a methodological process that strategically combines computational and human-driven coding methods can also lead to a construct map that could be used more seamlessly with later—perhaps larger scale—applications of ML, through supervised methods. Because the construct map was developed in part through the use of unsupervised ML methods (which utilized natural text), we saw the initial viability of the construct map as a framework for coding, especially when we coded a larger number of responses. While the degree of accuracy that we observed is lower than others have reported in supervised applications (e.g., Beggrow et al. 2014), it compares to most other uses of ML for science assessment (e.g., Table 1 from Gerard and Linn 2016 and Table A2 from Zhai et al. 2020). Also, importantly, because of the role of our human-driven analysis, the construct map is better aligned with the theorized aspects of generality with which we began and kept in mind as we interpreted the groupings of students' responses. For this reason, it may be more theoretically meaningful to those studying epistemic considerations.

Assessing Students' Considerations of the Generality of their Model-Based Explanations

Drawing upon scholarship and curriculum development efforts that emphasize involvement in knowledge-building practices (Ford and Forman 2006; Lehrer and Schauble 2006; Sandoval 2005; Schwarz et al. 2009), and the epistemic considerations undergirding their participation in those practices (Berland et al. 2016; Krist 2020; Manz 2015; Ryu and Sandoval 2012) we sought to center students' epistemic ideas as a key learning outcome. Most assessment efforts have focused explicitly on the structure of students' accounts, leading to the absence of the construct underlying the *how* and *why* aspects of students' involvement in science practices, especially with respect to the key consideration of generality.

Our approach to assessing students' epistemic ideas first relied upon eliciting students' rationales *about* their model-based explanations, then analyzing these rationales. Specifically, we used questions embedded in students' everyday classroom activities that asked them to reflect upon the generality (or specificity) of their model-based explanations. We then analyzed students' responses for their rationales for generality or specificity using a methodological approach that was responsive to the ideas students expressed through their written responses as well as theoretical aspects of generality. Our analysis resulted in a construct map that is both theoretically aligned and reflective of the empirical patterns in the data—some of which we did not anticipate.

Such a process is especially important given the nature of epistemic considerations are not a discrete idea or concept, but a constellation of ideas that students bring to the task of motivating their knowledge-building work. If a key goal for science is keeping content and practices coupled while also measuring something across grades and content areas (DeBarger et al. 2013), this construct map may serve as useful guidance for researchers interested in supporting students' meaningful participation in science knowledge–building practices. Similarly, embedded assessments containing comparable prompts may be useful to scholars seeking to study learning that spans more than one unit or content area. Moreover, the combination of embedded assessments and a CGT approach may be useful to other researchers studying complex constructs, particularly in light of calls for "a move away from reliance on the efficiency and affordability of multiple-choice items, and towards the use of more authentic tasks" (Haudek et al. 2019, p. 1; see also Laverty et al. 2016).

Limitations and Recommendations for Future Research

Some limitations of this work concern what we see as a trade-off between the goals of science assessment and ML: we analyzed responses in more than one content area. We recognize that studying multiple units may have led to a less-optimal supervised ML application and that more sophisticated data pre-processing (e.g., using more sophisticated stopword removal techniques) and/or validation efforts (e.g., examining for what content areas agreement between human and automated codes was better) could yield benefits.). Yet, we intended to develop a construct that could be used across science content areas and practices. Our results do present an effective proof of concept: although we developed these codes that do seem to be very specific to the task at hand (e.g., acetate is a key term for Cluster 4; see Table 3), when we added content areas (by coding all 1885 responses, in contrast to the 173 we used as a proof-of-concept), the construct map was able to be used effectively. These results suggest that conducting more nuanced (and time-intensive) analyses would be worthwhile to continue examining the use of ML techniques in analyzing students' epistemic ideas.

Another limitation concerns our modifications to the CGT method, particularly by analyzing a smaller corpus of student responses at first (and through the clustering approach, step 1, and the refinement of the construct map, step 2), after which we manually coded a larger number of student responses. In our future uses of the method, we would likely analyze the entire corpus of available student responses and revise and apply the construct map based on this larger dataset from the outset, and so we recommend for researchers to do the same—and to consider how the premise of CGT (that ML and human-driven interpretive coding can be integrated in useful ways) can be interpreted and applied creatively in other future applications. Last, while our data was longitudinal in nature, few students had data from all five time points. A longitudinal dataset with fewer missing responses may reveal even more nuance in the ways in which students' considered the epistemic criterion of generality.

Future research could also consider other types of data sources for use as part of an assessment approach similar to that used in this study. In particular, we highlight the potential of using ML techniques with audio and visual data as a highly complex, less-utilized (apart from in exclusively qualitative methods) source of data to capture an even more authentic and embedded version of students' participation in science practices.

Conclusion

Assessing students' involvement in science practices as emphasized in recent approaches in science education presents grand challenges to science education scholars and assessment experts (Pellegrino 2013). In this study, we sought to use ML for a novel purpose in science assessment: developing a construct map, which can be used to code students' written responses. We showed how ML and human-based methods can be integrated in a way that may have wide application, particularly as the broader science education community addresses the question of how to measure science learning in a way that does not minimize and simplify, but instead highlights the knowledge-building work of science learning.

Compliance with Ethical Standards

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee (Northwestern University #STU00034615 and Wright State University #FWA00002427) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Conflict of Interest The authors declare that they have no conflicts of interest.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

Allaire, J. J., & Chollet, F. (2019). *keras: R interface to 'Keras'*. R package version 2.2.5.0. https://CRAN.R-project.org/package=keras

Anderson, D. J., Rowley, B., Stegenga, S., Irvin, P. S., & Rosenberg, J. M. (2020). (advance online publication). Evaluating content-related validity evidence using a text-based, machine learning procedure. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12314.

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: how closely do they match clinical interview performance? *Journal of Science Education and Technology, 23*(1), 160–182.

Benoit, K., Chester, P., & Müller, S. (2019a). *quanteda.classifiers: models for supervised text classification*. R package version 0.1. http://github.com/quanteda/quanteda.svm

Benoit, K., Muhr, D., and Watanabe, K. (2019b). Stopwords: Multilingual Stopword lists. R package version 1.0. https://CRAN.R-project.org/package=stopwords

Bergman, L. R., & El-Khouri, B. M. (1999). Studying individual patterns of development using I-states as objects analysis (ISOA). *Biometrical Journal: Journal of Mathematical Methods in Biosciences, 41*(6), 753–770.

Berland, L., & Crucet, K. (2016). Epistemological trade-offs: accounting for context when evaluating epistemological sophistication of student engagement in scientific practices. *Science Education, 100*(1), 5–29.

Berland, L. K., Schwarz, C. V., Krist, C., Kenyon, L., Lo, A. S., & Reiser, B. J. (2016). Epistemologies in practice: making scientific practices meaningful for students. *Journal of Research in Science Teaching, 53*(7), 1082–1112.

Bouchet-Valat, M. (2014). SnowballC: snowball stemmers based on the C libstemmer UTF-8 library. R package version 0.5, 1.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education, 25*(1), 60–117.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: a theoretical framework for evaluating inquiry tasks. *Science Education, 86*(2), 175–218.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220.

Core Team, R. (2019). *A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing https://www.R-project.org.

DeBarger, A. H., Penuel, W. R., & Harris, C. J. (2013). Designing NGSS assessments to evaluate the efficacy of curriculum interventions. *Invitational Research Symposium on Science Assessment.* Washington, DC: K-12 Center at ETS. Retrieved from http://www.k12center. org/rsc/pdf/debarger-penuel-harris.pdf.

Dickes, A. C., Sengupta, P., Farris, A. V., & Basu, S. (2016). Development of mechanistic reasoning and multilevel explanations of ecology in third grade using agent-based models. *Science Education, 100*(4), 734–776.

Duncan, R. G., & Tseng, K. A. (2011). Designing project-based instruction to foster generative and mechanistic understandings in genetics. *Science Education, 95*(1), 21–56.

Ford, M. J. (2015). Educational implications of choosing "practice" to describe science in the next generation science standards. *Science Education, 99*(6), 1041–1048.

Ford, M. J., & Forman, E. A. (2006). Chapter 1: redefining disciplinary learning in classroom contexts. *Review of Research in Education, 30*(1), 1–32.

Fram, S. M. (2013). The constant comparative analysis method outside of grounded theory. *The Qualitative Report, 18*, 1.

Gerard, L. F., & Linn, M. C. (2016). Using automated scores of student essays to support teacher guidance in classroom inquiry. *Journal of Science Teacher Education, 27*(1), 111–129.

Giere, R. N. (1988). *Explaining science: a cognitive approach*. Chicago: University of Chicago Press.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: measuring students' science inquiry skills using educational data mining. *The Journal of the Learning Sciences, 22*(4), 521–563.

Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist, 50*(1), 43–57.

Gotwals, A. W., & Songer, N. B. (2013). Validity evidence for learning progression-based assessment items that fuse core disciplinary ideas and science practices. *Journal of Research in Science Teaching, 50*(5), 597–626.

Greene, D., Hoffmann, A. L., & Stark, L. (2019). *Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning*, *Hawaii international conference on system sciences (HICSS)*. HI: Maui.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice, 38*(2), 53–67.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

Haudek, K. C., Osborne, J., & Wilson, C. D. (2019). Using automated analysis to assess middle school students' competence with scientific argumentation. In *National Conference on Measurement in Education*. Toronto: NCME.

Helleputte, T. (2017). LiblineaR: linear predictive models based on the Liblinear C/C++ library. *R package version, 2*, 10–18.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261–266.

Inkinen, J., Klager, C., Juuti, K., Schneider, B., Salmela-Aro, K., Krajcik, J., & Lavonen, J. (2020). High school students' situational engagement associated with scientific practices in designed science learning situations. *Science Education, 104*(4), 667-692.

Jiménez-Aleixandre, M. P., Bugallo Rodríguez, A., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": argument in high school genetics. *Science Education, 84*(6), 757–792.

Kelly, G. J. (2008). Inquiry, activity and epistemic practice. In R. A. Duschl & R. E. Grandy (Eds.), *Teaching Scientific Inquiry* (pp. 99–117). https://doi.org/10.1163/9789460911453_009.

Kolodner, J. L. (Ed.). (1993). *Case-based learning*. Dordrecht: Kluwer Academic Publishers.

Krajcik, J., McNeill, K. L., & Reiser, B. J. (2008). Learning-goals-driven design model: developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Science Education, 92*(1), 1–32.

Krajcik, J., Reiser, B., Sutherland, L., & Fortus, D. (2011). *IQWST: investigating and questioning our world through science and technology* (middle school science curriculum materials). Greenwich: Sangari Active Science.

Krist, C. (2020). Examining how classroom communities developed practice-based epistemologies for science through analysis of longitudinal video data. *Journal of Education & Psychology, 112*(3), 420–443. https://doi.org/10.1037/edu0000417.

Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science, 9*(5), 178–181.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*(2), 363–374.

Laverty, J. T., Underwood, S. M., Matz, R. L., Posey, L. A., Carmel, J. H., Caballero, M. D., Fata-Hartley, C. L., Ebert-May, D., Jardeleza, S. E., & Cooper, M. M. (2016). Characterizing college science assessments: the three-dimensional learning assessment protocol. *PLoS One, 11*(9), e0162333.

Lead States, N. G. S. S. (2013). *Next generation science standards: for states, by states*. Washington, DC: National Academies Press.

Lehrer, R., & Schauble, L. (2006). Cultivating model-based reasoning in science education. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (p. 371–387). Cambridge University Press.

Lehrer, R. & Schauble, L. (2015). Developing scientific thinking. In L. S. Liben & U. Müller (Eds.), *Cognitive processes.* Handbook of child psychology and developmental science (Vol. 2, 7th ed., pp. 671-174). Hoboken, NJ: Wiley.

Lehrer, R., Schauble, Leona, & Petrosino, A. J. (2001). Reconsidering the role of experiment in science education. *Designing for science: implications from everyday, classroom, and professional settings*, 251–278.

Manz, E. (2012). Understanding the codevelopment of modeling practice and ecological knowledge. *Science Education, 96*(6), 1071–1105.

Manz, E. (2015). Representing student argumentation as functionally emergent from scientific activity. *Review of Educational Research, 85*(4), 553–590.

McNeill, K., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *The Journal of the Learning Sciences, 15*(2), 153–191.

Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to develop a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching, 54*(8), 1024–1048.

National Research Council. (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

National Research Council (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press.https://doi.org/10.17226/18409.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. *Journal of Science Education and Technology, 21*(1), 183–196.

Nelson, L. K. (2020). Computational grounded theory: a methodological framework. *Sociological Methods & Research, 49*(1), 3–42. https://doi.org/10.1177/0049124117729703.

Passmore, C., Schwarz, C. V., & Mankowski, J. (2017). Developing and using models. In C. V. Schwarz, C. Passmore, & B. J. Reiser (Eds.), *Helping students make sense of the world using next generation science and engineering practices* (pp. 109–135). Arlington, VA: NSTA Press.

Pei, B., Xing, W., & Lee, H. S. (2019). Using automatic image processing to analyze visual artifacts created by students in scientific argumentation. *British Journal of Educational Technology, 50*(6), 3391–3404.

Pellegrino, J. W. (2013). Proficiency in science: assessment challenges and opportunities. *Science, 340*(6130), 320–323.

Penuel, W. R., Turner, M. L., Jacobs, J. K., Van Horne, K., & Sumner, T. (2019). Developing tasks to assess phenomenon-based science learning: challenges and lessons learned from building proximal transfer tasks. *Science Education, 103*(6), 1367–1395.

Popper, K. R. (1959). The propensity interpretation of probability. *The British Journal for the Philosophy of Science, 10*(37), 25–42.

Reiser, B. J., Kim, J., Toyama, Y., & Draney, K. (2016). Multi-year growth in mechanistic reasoning across units in biology, chemistry, and physics. *Paper presented at NARST, April, 14*, 2016.

Rosenberg, J. M., & Lishinski, A. (2018). *clustRcompaR: easy interface for clustering a set of documents and exploring group-based patterns* [R package]. https://github.com/alishinski/clustRcompaR

Ryu, S., & Sandoval, W. A. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. *Science Education, 96*(3), 488–526.

Saldaña, J. (2016). *The coding manual for qualitative researchers*. Sage.

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education, 89*(4), 634–656.

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction, 23*(1), 23–55.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L. O., Archer, A., Fortus, D., & Krajcik, J. (2009). Developing a learning progression for scientific modeling: making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 46*(6), 632–654.

Schwarz, C. V., Passmore, C., & Reiser, B. J. (2017). *Helping students make sense of the world using next generation science and engineering practices*. Arlington, VA: NSTA Press.

Shaffer, D. W. (2017). *Quantitative ethnography*. Madison: Cathcart Press.

Sherin, B. (2013). A computational study of commonsense science: an exploration in the automated analysis of clinical interview data. *The Journal of the Learning Sciences, 22*(4), 600–638.

Shin, N., Stevens, S. Y., & Krajcik, J. (2010). Tracking student learning over time using construct-centred design. In *Using Analytical Frameworks for Classroom Research* (pp. 56–76). Routledge.

Tabak, I., & Reiser, B.J. (1999). Steering the course of dialogue in inquiry-based science. Paper presented at the *Annual Meeting of the American Educational Research Association Montreal*, Canada.

Thagard, P. R. (1978). The best explanation: criteria for theory choice. *Journal of Philosophy, 75*(2), 76–92.

Wiley, J., Hastings, P., Blaum, D., Jaeger, A. J., Hughes, S., Wallace, P., Griffin, T. D., & Britt, M. A. (2017). Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science. *International Journal of Artificial Intelligence in Education, 27*(4), 758–790.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. London: Routledge.

Zangori, L., Forbes, C. T., & Schwarz, C. V. (2015). Exploring the effect of embedded scaffolding within curricular tasks on third-grade students' model-based explanations about hydrologic cycling. *Science & Education, 24*(7–8), 957–981.

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement, 76*(2), 280–303.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education, 56*(1), 111–151.