JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc Article

Developing Computer Resources to Automate Analysis of Students' Explanations of London Dispersion Forces

Keenan Noyes,* Robert L. McKay, Matthew Neumann, Kevin C. Haudek, and Melanie M. Cooper

Cite This: *J. Chem. Educ.* 2020, 97, 3923–3936 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Computer-assisted analysis of students' written responses to questions is becoming a possibility due to developments in technology. This could make such constructed response questions more feasible for use in large classrooms where multiple choice assessments are often considered a more practical option. In this study, we use a previously developed prompt and coding scheme to characterize students' explanations of the origins of London dispersion forces in order to develop machine learning resources that can carry out such an analysis for large numbers of students. We found that by using large numbers of human coded student responses (N = 1,730) we could subsequently automatically characterize students' responses at a high level of accuracy compared to human coders. Furthermore, these resources were developed using responses from several different groups of students across multiple institutions to ensure both that our resources can work well with students from different backgrounds and that these computer resources can detect the different ways in which students explain this phenomenon. Such resources may help instructors to administer more complex open-ended assessment tasks to larger numbers of students and analyze the responses capturing language corresponding to causal mechanistic reasoning. Instructors could then use this information to better support their students' learning.

KEYWORDS: *First-Year Undergraduate/General, Chemical Education Research, Testing/Assessment, Noncovalent Interactions*

FEATURE: Chemical Education Research

INTRODUCTION

With our rapid advances in technology, automating assessment of students' responses has become an emerging possibility for the field of education. However, the use of automated text analysis technology is relatively new in the field of chemistry education research. While there are many assessment systems that integrate forced choice or numerical response tasks, the assessment of student constructed written responses to complex chemistry prompts is not as available. Such assessment tasks can provide instructors with important information about what students know and can do; however, there are currently limited resources that allow for meaningful analysis of student explanations of chemical phenomena. Here, we use our work on students' explanations of the origins of London dispersion forces (LDFs) to explore an approach to the development of machine learning resources. We also investigate whether these resources can code undergraduate students' responses similarly to how humans code those responses, and whether these resources can detect different signals (i.e., different ways in which students explain this phenomenon).

Importance of Assessments

Assessment of student learning can be thought of in terms of evidence-based arguments.[1,2] Using this framework, evidence (in the form of student responses) is gathered from assessment items and subsequently used to support arguments about what students know and can do. Once instructors have such evidence, they can use it to evaluate the depth of student learning and to revise learning materials and instructional methods to support more robust understanding. In this way, assessments become more than just a way to evaluate students, but also an important tool to support students' learning by providing the instructor important feedback about the design and effectiveness of educational materials and teaching strategies.[3]

However, this design cycle is only effective if the assessment tasks elicit appropriate evidence about student learning. While instructors may infer that students are using appropriate reasoning to answer a question, if that question does not explicitly elicit such evidence, there is a strong likelihood that some students are using rules of thumb and learned or taught heuristics to answer questions.[4,5] For example, many students who can answer multiple choice questions about intermolecular forces (IMFs) and their role in physical properties have been shown to construct drawn representations of IMFs acting within a molecule (rather than between molecules).[6] It has been noted that students may use the presence of hydrogen

Received: May 8, 2020
Revised: September 21, 2020
Published: October 19, 2020

bonding to predict relative boiling points, while at the same time having erroneous ideas about the nature of hydrogen bonding itself.[5,7,8] Indeed, the idea that boiling water produces hydrogen and oxygen becomes less surprising when we recall that students are told that boiling water breaks hydrogen bonds. The fact that there are many students who go through years of chemistry instruction without understanding the nature of IMFs, and the fact that their instructors were unaware of this problem, may be because many traditional assessments do not elicit explicit evidence of the students' understanding of IMFs.

This process is even more challenging because the design of the task and associated coding scheme is crucial to whether appropriate reasoning can be elicited and captured. In this study, we characterized students' responses on the basis of a previous rubric we developed that captures the degree to which a student's explanation provides a causal mechanism for the phenomenon of LDFs.[9,10] This type of analysis is important because supporting causal mechanistic thinking is an important goal of science education.[11] Developing an understanding of how and why unseen entities behave and give rise to phenomena gives learners the ability to generate robust explanations and make predictions.[12] This approach goes beyond simply identifying what problematic ideas a student might have. Instead, by identifying the resources students use to explain the mechanism, we can begin to understand how best to support student learning.

Designing Assessments

While there is certainly a place for multiple choice assessments, such items typically do not provide the kind of evidence that would allow us to make convincing arguments about what students know and can do, particularly if we want to go beyond recognition of fragmentary information or algorithmic problem solving. Whether a student gets the answer right or wrong, their thinking is not visible, so we cannot know how the student arrived at that answer or the type of reasoning they have engaged in. An ideal way to explore student thinking would be to ask them directly, for example, in an interview setting. However, such a method becomes impractical in classroom settings where an instructor may be responsible for several hundred students. Compromises must be made, so instructors instead often use multiple choice or constructed response assessments (in which the student must generate a text response). However, these two types of assessments do not necessarily elicit the same student understanding. For example, Nehm and Schonfeld[13] used multiple choice, constructed response, and interview questions to assess their students' understanding of natural selection. They found that the understanding conveyed in the interview responses aligned more closely with the constructed responses compared to the multiple choice responses. Other studies have reported similar findings, that multiple choice questions tend to overestimate what students know compared to constructed response questions.[14,15] Additionally, answering only multiple choice questions or questions that do not elicit reasoning does not explicitly provide students with the opportunity to reason and reflect on their responses. This process is integral to learning; indeed, asking deep explanatory questions is one of the few pedagogical techniques to promote learning that is supported by strong evidence.[16]

However, asking such questions typically requires multiple rounds of prompt design, hand coding of responses, and refinement of the prompts from the resulting evidence in order to better elicit student ideas. For example, the constructed response questions used by Nehm and Schonfeld[13] were originally developed by Bishop and Anderson[17] and were subjected to multiple rounds of design, coding, and redesign. We have previously reported on our efforts to elicit and characterize stronger evidence about IMFs, acid−base chemistry, simple nucleophilic substitutions, and LDFs.[6,9,18−20] The assessments in each of these studies went through a similar iterative design process. Assessment items that are too vague typically do not produce rich responses, while prompts that are too specific tend to signal to students what the desired responses should be. Prompts that do not provide enough direction about what is required, or that do not activate appropriate resources may lead students who might otherwise provide a rich response to give a more simplistic answer. The goal is to find the prompt that is "just right", allowing students to tell us what they know.

For example, consider the prompt we used in this study: our previously developed LDF prompt.[9,10] We designed this question to explore students' explanations of how and why neutral atoms attract. We wanted to know if students could leverage their knowledge of the electrons and protons within the atoms and unpack their properties (e.g., their charges, how they move) to explain this phenomenon. Such an explanation that links behaviors of the entities a scalar level below the phenomenon would be evidence of causal mechanistic reasoning, a powerful form of reasoning in science.[12] Our initial efforts to elicit this kind of response resulted only in surface-level descriptions. In order to develop a question that elicited causal mechanistic reasoning and a subsequent coding scheme to characterize those responses we interviewed students, piloted multiple prompts, and analyzed hundreds of students' responses. For more information about this process, see Becker et al.[9] and Noyes and Cooper.[10]

Role of Formative, or Low-Stakes, Assessments

Formative assessment typically refers to assessments that are low-stakes, that provide students with an opportunity to "try out" ideas without penalty, and to receive feedback, ideally leading to improvement over time.[21,22] Generally, and particularly for large enrollment courses, it is not feasible to use such items for assessments that must be responded to on a daily basis. Indeed, this is one reason why many large enrollment courses have defaulted to machine scorable multiple choice and fill in the blank items that can be easily scored and even provided with automated feedback. In our work, we do use items that require explanations, arguments, and drawings. In these assessments, students are awarded credit for completion, not for accuracy; that is, we do not grade the answers but instead provide aggregate feedback and discussion in larger groups. However, if we want to learn how students are meeting the challenges of such items, it becomes necessary to hand score which is time-consuming and expensive both in terms of resources and personnel. Generally, we have found that students' responses on these formative assessments are quite similar to their responses on analogous summative constructed response items.[10]

Case for Machine Learning

Machine learning may provide an answer to this conundrum by automating the analysis of the students' responses to such items. There are growing numbers of reports describing such approaches to the analysis of short, concept-based, text

https://dx.doi.org/10.1021/acs.jchemed.0c00445
J. Chem. Educ. 2020, 97, 3923−3936

responses in a variety of STEM education areas including acid−base chemistry,[23−25] the definition of randomness in statistics,[26] the role of the stop codon in genetics,[27] and biological mechanisms of weight loss,[28] among others. We have been working with a machine learning group, the Automated Assessment of Constructed Response (AACR) group,[29,30] to develop machine learning resources capable of analyzing the LDF question that we previously developed.[9,10] The AACR group has developed tools available on the web, such as the Constructed Response Classifier (CRC) tool which is discussed in this paper, that use open-source machine learning techniques to create computer models that are capable of mimicking human analyses by identifying key disciplinary concepts or reasoning patterns in student responses. Once these computer models have been developed, they can be used to analyze large numbers of new text responses from students just as an expert would, but requiring only a few minutes. A tool like this would then provide instructors a way to incorporate constructed response questions more routinely in their teaching. For reasons we will discuss later, such resources are not intended to give high-stakes feedback for individual students but, rather, useful and timely information about how a set of students understands or reasons with a particular idea.

Overview of the Constructed Response Classifier

One of the biggest challenges with automating analysis of student responses is developing computer resources that perform well and that are able to recognize nuanced and complex explanations in short text responses. AACR's CRC is a promising tool for machine learning analysis because (1) nearly all of the process is automated and (2) this technology uses several machine learning algorithms in an ensemble.[31] The first point is important because while other software, such as SPSS Modeler, can perform lexical analysis, much more human input is needed. For example, Dood et al. used SPSS Modeler for automated analysis of student responses, and while the software could identify common terms in the responses, humans needed to specify synonyms and create categories and rules within the software in order to develop the predictive models.[23,24] This means that the researchers need to develop a meaningful way to analyze the responses, code the responses, and then translate their subsequent scheme into the software manually.

In contrast, the CRC contains a set of open-source, supervised machine learning classification algorithms developed by Jurka et al.[32] which can be used to predict scores of responses after "learning" from a set of previously scored student responses. As part of this routine, text in student responses is automatically extracted and parsed into *n*-grams, words and sets of words up to a defined number *n*, which are used as independent variables in the classification algorithms. Such an approach reduces the need for developing lexical resources like defining term dictionaries or creating combinatorial rules (see Nehm et al. and Kaplan et al.).[26,33] That is, the researcher only needs to develop and apply a coding scheme; the machine learning tools handle the rest thereby reducing the time and effort needed from researchers to develop these resources.

Additionally, the classification procedure developed by Jurka et al.[32] combines results from 8 machine learning algorithms to predict a single overall score for each response. By using multiple classification algorithms, the resulting predictions are generally more accurate than using only one algorithm (e.g., Optiz and Maclin).[34] The outcome of this process is that, once trained with human coded responses, the CRC is able to analyze "raw" student responses and predict codes on the new set of data.

Challenges with Machine Learning and Potential Solutions

Even with the machine learning techniques that the CRC uses, the fundamental challenges of automated analysis remain. That is, these computer models are mimicking human analysis of text responses, and therefore, the performance of the model is highly dependent on the rubric and the text responses used to train the model. In developing our computer resources, we tried to address both of those dimensions.

The first dimension is the rubric used to code the responses. Since the computers are mimicking the human analysis, if the humans cannot code reliably, neither will the machine. The rubric we developed to characterize students' LDF explanations was the product of analyzing student interviews, homework assignments, and exam responses, from which we iteratively refined the rubric to capture all the ways students could explain this phenomenon using causal mechanistic reasoning. The end result is a rubric that can capture reasoning and can be used reliably.

The second dimension is that of the text responses. Automated analysis techniques work by identifying patterns in the words and terms used by responses classified into each coding category so that, when presented with a new response, the model can analyze the words in the response to make a prediction. The better the model can identify the key features of each coding category (based on the patterns of words in the response), the more likely the computer is to score that response like the human. This requires a large number of human coded responses (in each coding category) upon which the computer model can be built. Therefore, we collected and took the time to code a very large number of student responses so critical components of each coding category would become salient. In our original analyses,[9] we defined six coding categories for student responses, but in subsequent work, we consolidated these into three coding categories.[10] As we discussed in our previous work,[10] we consolidated the number of coding categories to capture causal mechanistic explanations more concisely and also create a more practical scheme for humans to use to analyze thousands of responses. This also assisted the development of automated resources; by decreasing the overall number of coding categories, we increased the number of responses in each coding category.

It is not just the number of responses that is important but also their content. If we want to develop resources that can be useful across different institutions and courses, then the model must be trained with responses from a variety of learning environments. Otherwise, if these resources encounter a response that explains the phenomenon causal mechanistically, but using words not captured in the training set, the model would be unable to code the response accurately. For example, Ha et al.[35] explored how accurately a model developed from responses collected at one institution could code responses from another institution. They found that, while the computer could accurately code some key concepts related to evolutionary change for students at both institutions, it did not accurately code all of the key concepts that they intended to capture. Ha et al.[35] note that one of the issues affecting the accuracy of the models across the different institutions is

https://dx.doi.org/10.1021/acs.jchemed.0c00445
J. Chem. Educ. 2020, 97, 3923−3936

difference in language patterns unique to each institution. Our study is similar to that of Ha et al.[35] in that we also collect responses from multiple undergraduate institutions and different student populations to get a wider range of the ways students explain our phenomenon under study. However, we build upon their findings in this study by developing, and subsequently testing, a model from a combination of responses from all the institutions to capture any differences in the lexical patterns of the responses. Collecting responses across different institutions to develop and test our resources is also important so that the technologies we develop are equitable; that is, these resources are able to characterize the responses from students of all different backgrounds like a human coder would. By addressing these challenges, we hope to develop machine learning resources that can code student reasoning as part of their LDF explanations as well as humans can. Such a resource could be used to give instructors meaningful feedback about how their students understand the mechanism by which this IMF operates to better support student learning.

RESEARCH QUESTIONS

1. How does the machine coding for causal mechanistic explanations compare to humans?
2. How does the machine coding for different groups of students compare?

METHODS

Strategy for Developing Machine Learning Resources

In the Methods section, we describe the process by which we collected and analyzed data to develop supervised machine learning models. Our strategy was to first collect and human code many student responses, from a variety of contexts. Then, these coded responses were used to train the machine learning algorithms in order to develop resources that are capable of coding new responses in the same way as a human coder. The accuracy of the developed models was tested by applying the computer model to new sets of student responses and then comparing the results to the human coding of a subset of these new responses.

Participants

Responses from four groups of students from three different undergraduate institutions are analyzed here. The data were collected in accordance with each institution's IRB protocols, and at each institution, students consented to having their responses used for research purposes. Students' responses were deidentified before analysis. In the sections below, we provide some additional context for each of the four groups of students. We present more thorough descriptions of the demographics of each group in Supporting Information Section S1. We note that the demographic information reported was determined by the Registrar's office from each institution and therefore has limitations (e.g., the conflation of a student's racial and ethnic identity, the reporting of only male and female gender identities).

Group 1A. Group 1A is composed of students in the first semester of general chemistry from a large midwestern public research institution (which we call institution 1 in this paper). Specifically, these students were taking the first part of a 2 semester transformed general chemistry curriculum CLUE (Chemistry, Life, the Universe and Everything)[36] in the fall semester. While General Chemistry 1 is also offered in the spring, a majority of students at this institution take General Chemistry 1 in the fall. Of the 2,497 students enrolled in this course in fall 2015, 91% responded to our activity and consented for their response to be used for research (N = 2,284). We used 950 of those responses for computer model training. We discuss more about these responses and how they were selected in the later section in this paper about model development. We analyzed more responses from this group than any other because the initial model development process requires lots of responses and we collected the most responses from this group. This group of 950 students was primarily White (69%), and about half of the group was female (55%) (Supporting Information Section S1).

Group 1B. We also collected responses from General Chemistry 1 students at institution 1 in the spring of 2016. In this paper, we describe these "off-sequence" students as group 1B. Some students take such sections because the university has required them to complete additional math courses to prepare for general chemistry. This may be why the difference in mean ACT math scores for group 1A (mean = 25.8, N = 1,925) and group 1B (mean = 24.5, N = 776) is statistically significant (see Supporting Information Section S1). Like group 1A, the students in group 1B also had the CLUE general chemistry curriculum.

Of the 1,070 students enrolled in this course, 86% of the students (N = 915) responded to our prompt and consented for their responses to be used for research. Of those 915, we randomly selected 350 responses (using a random number generator) for this study. We selected only a portion of the total number of responses for machine learning development and testing because we wanted to have a group similar in size to groups 2 and 3. Like group 1A, the 350 randomly selected group 1B students were primarily White (70%), and about half were female (54%) (Supporting Information Section S1).

Group 2. Students in group 2 were enrolled at another large midwestern public research university, which we call institution 2 in this paper. We recruited this group from the first semester of a 2 semester general chemistry course in spring of 2018. Unlike institution 1, this institution used a traditional curriculum. The textbook listed on the course syllabus was *Chemistry: The Central Science* (14th ed.) by Theodore Brown.[37] We note that in this curriculum students received less explicit instruction about the construction of causal mechanistic explanations than those students in the CLUE curriculum, where such reasoning is an explicit focus of the course.

Of the 721 students enrolled, 53.3% of students (N = 384) responded and consented for their responses to be used for research purposes. The majority of those who responded to the activity were female (65%). While we did not collect information about the racial/ethnic identities of the students in this group (we did not get IRB approval to collect this information from this institution), we can approximate the demographic makeup of this group from the information available on the registrar's website about the entire institution. For the entire undergraduate student body (N = 23,856), the students were primarily White (73%) (Supporting Information Section S1).

Group 3. Group 3 is made up of students from institution 3, a large southeastern public research university, who had the first semester of general chemistry in either fall 2016 or fall 2017. This general chemistry course also used the CLUE curriculum but in a different format from institution 1. At

https://dx.doi.org/10.1021/acs.jchemed.0c00445
J. Chem. Educ. 2020, 97, 3923−3936

institution 1, the course was taught in large lecture sections (between 360 and 450 students per section) with smaller weekly recitations, while at institution 3 the course was taught in a partially flipped classroom format, where 100−200 students spent the majority of the time during class working in groups, receiving instruction and additional activities as homework outside of the classroom.[38]

Of the 449 students who received the activity over the two years, 77.1% of students (N = 346) submitted responses and consented for their work to be used for research purposes. Of these students, the majority were Hispanic (71%), and 8% were White (Supporting Information Section S1). This is quite different from the racial/ethnic identities of the students from institutions 1 and 2 where no more than 10% of the students identified as Hispanic. Additionally, just over half of this group was female (52%).

Question Prompt

To probe students' understanding of how LDFs arise, we asked students to explain why two helium atoms attract one another. This prompt was developed as part of our previous work eliciting causal mechanistic explanations of LDFs.[9,10] Originally, we developed this prompt to include a corresponding drawing component, but for the purposes of this study, we focus only on automatically analyzing the text portion. While groups 1A, 1B, and 3 still received that drawing component, it was administered on a separate slide after the text prompt. When the student reached the drawing prompt, they could not return to the text prompt to change their response. Practically, this means that the student answered the text prompt without ever seeing the drawing prompt; therefore, it appears that we are able to code students' text responses independently from their drawings.

For groups 1A, 1B, and 3, this prompt (Figure 1) was included as part of their homework activities. These formative assessments were a required part of the course, but the students received course credit for completing the activity, not for the correctness of their response. Both institutions 1 and 3 used the online beSocratic system[39] for homework which allowed us to collect responses digitally, a helpful feature when automating analysis of the responses. Additionally, since this question was included in their homework, the students completed this activity shortly after instruction of LDFs. In the CLUE curriculum, introduction of IMFs including LDFs occurs early as these ideas are continually built upon throughout the course.

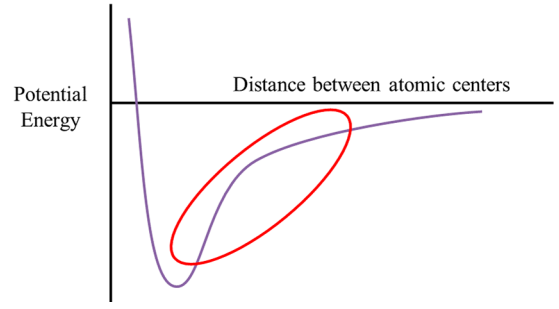

Figure 1. LDF prompt.

For group 2, we administered the activity in a different manner. In this course, the instructor gave the students the opportunity to answer this question for a small amount of extra credit, but it was not a required part of the course. This may explain why the response rate was lower for institution 2 compared to institutions 1 and 3 where the question was part of a required homework activity. Since it was not part of a broader activity, the prompt was asked as a standalone question. Additionally, the instructor did not have access to beSocratic and so instead used the online survey system Qualtrics[40] to collect the students' text responses (we did not administer the drawing component to this group). The timing for this activity also differed as it was given to students near the end of the semester. We acknowledge that these conditions differed substantially from those at institutions 1 and 3, but the purpose of this study is to collect a lexically diverse set of responses to train and test machine learning resources, not to compare across groups. By giving this prompt in contexts where the administration of the prompt, student demographics, and instructional environment differ, we accomplished this goal.

Coding Scheme

To analyze the responses, we used the coding scheme that we had previously developed to characterize the degree to which the students engaged with causal mechanistic reasoning.[10] This holistic, mutually exclusive coding scheme places responses into one of three categories: nonelectrostatic (NE), electrostatic causal (EC), and causal mechanistic (CM) (Table 1). NE responses fail to provide any electrostatic evidence for this interaction. Electrostatic causal responses discuss the role of electrostatic attractions in this interaction but do not include the mechanism by which these interactions form. Causal mechanistic responses go further, explaining the happenings at the scalar level below: how the electrons temporarily can localize on one side of atom which results in the separation of charge that causes this interaction.

Human Coding of Responses

Supervised machine learning relies on a set of "labeled" data in order to train or develop the computer model. As used here, the human codes assigned to students' responses were the "labels" necessary for training the computer model. We report descriptive information about the response length and provide some additional example responses in Supporting Information Section S2.

Before conducting any coding, we deidentified all student responses and established inter-rater reliability (IRR) between the researchers. In this process, two people independently coded small sets of students' responses and then calculated Cohen's kappa, a measure of agreement that also considers the probability of agreement by chance.[41] Once the Cohen's kappa value surpassed 0.7, a value corresponding to a "substantial" level of agreement,[42] we determined that we had reached IRR. At this point, we could begin coding the responses that would later be used to train and test the automated resources. We provide an overview of the process by which we established IRR and conducted subsequent coding in the following paragraphs and Figure 2.

Before coding the group 1A responses, author K.N. and an undergraduate researcher conducted IRR. After eight rounds of coding (30 responses in rounds 1−7 and 50 responses in

https://dx.doi.org/10.1021/acs.jchemed.0c00445
J. Chem. Educ. 2020, 97, 3923−3936

Table 1. Overview of Coding Scheme from Noyes and Cooper[10]

Type of Response	Text Features	Student Example
Nonelectrostatic (NE) text response	The response does not include reasonable electrostatic evidence of the interaction. Instead, the response provides nonelectrostatic evidence or does not address the intermolecular interaction between molecules.	"The two atoms are attracted because of the electromagnetic forces that exist between them"
Electrostatic causal (EC) text response	The response indicates that electrostatic charges cause the interaction. Examples of electrostatic causal evidence include subatomic particles, overall charge of the atom, partial charges, etc. These responses do not include a mechanism by which a separation of charge occurs.	"Atoms attract to each other [because] the [London] dispersion forces make the partially negative end of one atom attract to the partially positive end of another atom."
Causal mechanistic (CM) text response	The response indicates that the interaction occurs due to electrostatic charges and includes the mechanism by which the instantaneous and/or the induced dipole forms.	"As the 2 atoms approach each other one of them becomes [instantaneously] dipole due to fluctuation in its electron cloud as most of the electrons are concentrated to one side of the atom making it have a partially positive and negative charges on opposite sides. This partially positive side of the attracts the [electrons] of the other atom making its electron cloud fluctuate and form an instantaneous dipole where both of them attract each other."

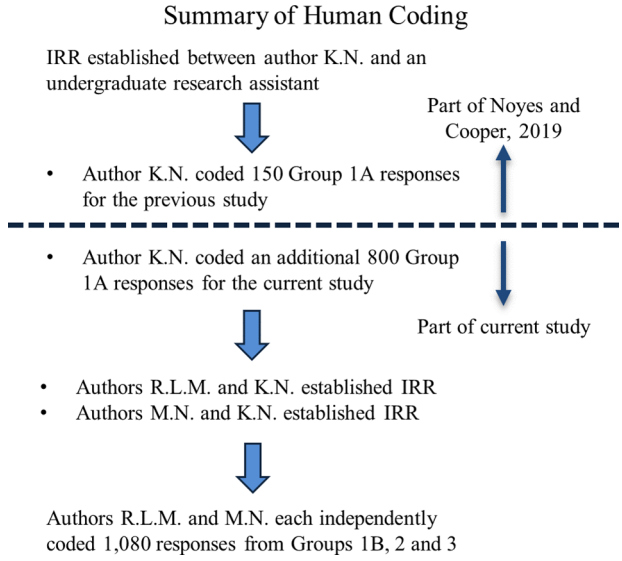

Figure 2. Summary of the human coding reported in this study with the distinction as to which data was reported in our previous study, Noyes and Cooper.[10]

round 8), we established IRR (Cohen's kappa = 0.78, 88% agreement). During this process, the two researchers met after each round of coding to discuss disagreements and, if needed, refine the coding categories. This process is discussed more in depth in the Supporting Information S1 of Noyes and Cooper.[10] Author K.N. then coded 950 responses from group 1A to train the initial computer model. We note that Author K.N. coded 150 of these responses as part of a previous study.[10]

The coding of responses from groups 1B, 2, and 3 was carried out by authors R.L.M. and M.N. Before coding, both authors separately conducted IRR with author K.N. using sets of 40 previously uncoded responses from group 1A. Authors R.L.M. and K.N. reached IRR after two rounds (Cohen's kappa = 0.88), and authors M.N. and K.N. also reached IRR after two rounds (Cohen's kappa = 0.81), with both Cohen's kappa values corresponding to "Almost Perfect" agreement.[42] While there are other statistical tests to calculate agreement between three coders such as the weighted Cohen's kappa, authors R.L.M. and M.N. began working on this project at different times and therefore were not trained on this coding scheme simultaneously.

With IRR established, both authors R.L.M. and M.N. coded the selected responses from groups 1B, 2, and 3. To minimize bias, the authors were not aware of the institution affiliation of the responses they were coding. These responses were coded individually in batches of approximately 100 responses, and then compared. When comparing, the two authors discussed any discrepancies between their individual codes and assigned a final, mutually agreed upon code to any disputed responses. This continued until the responses of an entire group were coded. This process was then repeated for all three groups. We characterized the initial level of agreement between the coding of authors R.L.M. and M.N. for each group by calculating Cohen's kappa and percent agreement values for their initial codes (Table 2).

General Overview of How the CRC Works

The AACR group has developed machine learning tools to automate the analysis of students' written responses. We used

https://dx.doi.org/10.1021/acs.jchemed.0c00445
J. Chem. Educ. 2020, 97, 3923−3936

Table 2. Initial Level of Agreement for Each of Three Cohorts between Authors R.L.M. and M.N.

Group	Number of Students	Cohen's Kappa	Percent Agreement
1B	350	0.71	81%
2	384	0.74	89%
3	346	0.69	81%

AACR's CRC web application (the developed model is now accessible through the AACR project website: beyondmultiplechoice.org) to develop resources to mimic our human coding of the responses to our LDF prompt. This app uses a series of eight machine learning algorithms derived from the open-source statistical package RTextTools developed by Jurka et al.[32] to predict the code for a response based on human coded responses (see also Sieke et al.[31]). Jurka et al.[43] provide a more detailed description of this package and its function, but we briefly describe the inner workings of the CRC tool here.

Before the training responses are input into RTextTools, the CRC "cleans" the responses, based on options selected by the user. In this process, the responses undergo "stemming" so that the suffixes (e.g., "attraction" becomes "attract"), stop words (e.g., "and", "the", "a", "in"), and numerical characters are removed. This leaves the important terms for the lexical analysis. The cleaned responses are then loaded into RTextTools.

Initially, the set of training responses is used to create a document-term matrix.[44] A document-term matrix parses out all of the individual words (unigrams) and pairs of words (bigrams) for each of the responses and captures it in a matrix. In Figure 3 we present a hypothetical document-term matrix for a hypothetical pair of simple student responses. With hundreds of longer student responses, this matrix can get very large, very quickly.

To train the computer model, machine learning algorithms use the document-term matrix and the corresponding human scores for each of the responses in the training set to generate a predictive model capable of processing a new response. We show an overview of this process with some hypothetical responses in Figure 3. Each algorithm uses the patterns of the presence and absence of all the unigrams and bigrams for each of the responses in the training set. This means that the algorithms are not simply looking for a list of predefined keywords but, instead, are identifying patterns based on all of the words (*n*-grams) in the response (as captured in the document-term matrix). We note that the training of each predictive model is fully automated: There is no human input at this step. To maximize the benefits of the range of machine learning techniques available and to minimize the downsides of any one machine learning algorithm, eight different algorithms are used to construct eight different predictive models: support vector machines,[45] supervised latent dirichlet location,[46] logitboost,[47] classification trees,[48] bagging classification trees,[49] random forests,[50] penalized generalized linear models,[51] and maximum entropy models.[52] When presented with a set of responses, the CRC cleans the responses, and then, RTextTools generates a document-term matrix for each particular response. All eight models then use the presence or absence of the unigrams and bigrams in the document-term

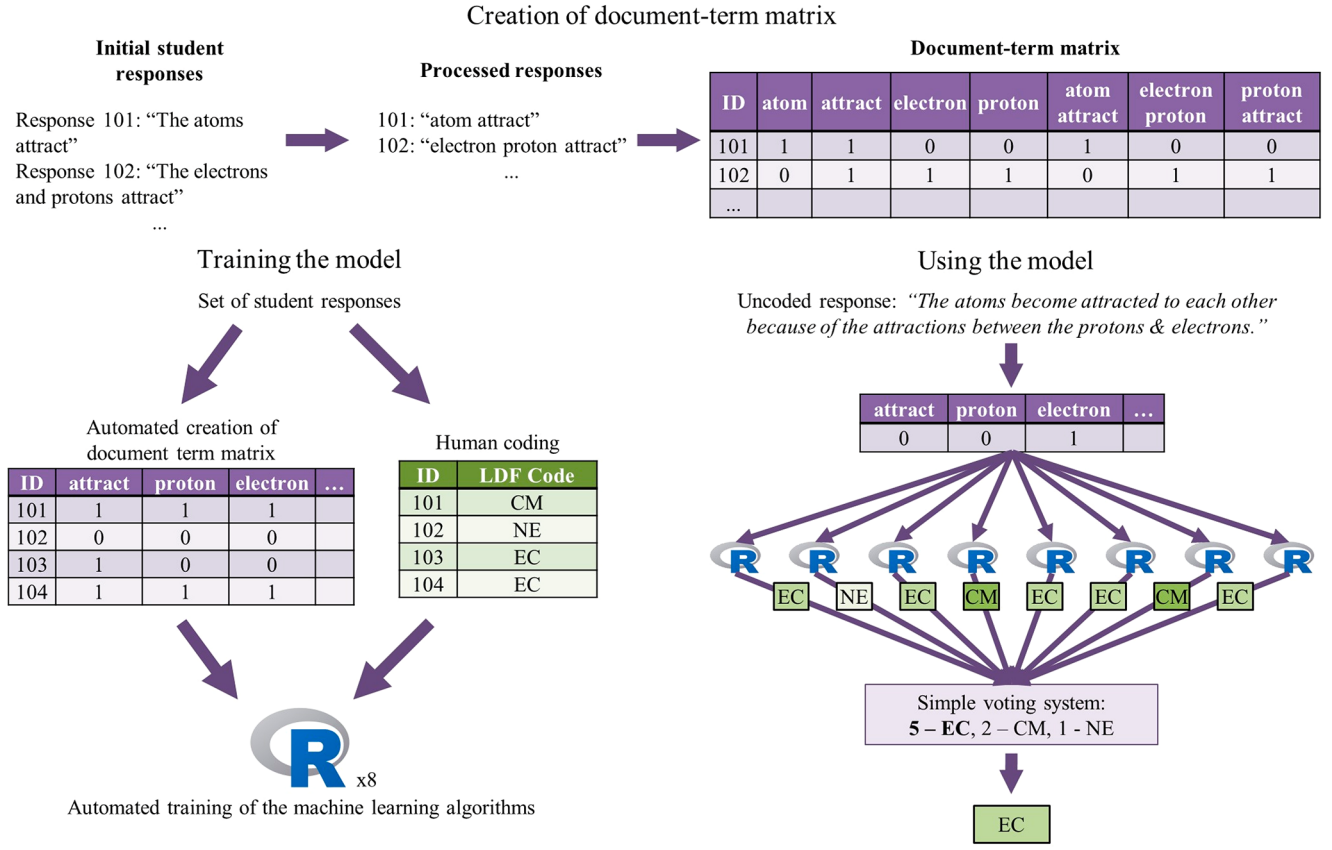

Figure 3. General overview of the automated coding process using several hypothetical student responses highlighting the creation of a document-term matrix, the training of the computer model, and the process by which the model then codes new responses.

https://dx.doi.org/10.1021/acs.jchemed.0c00445
J. Chem. Educ. 2020, 97, 3923−3936

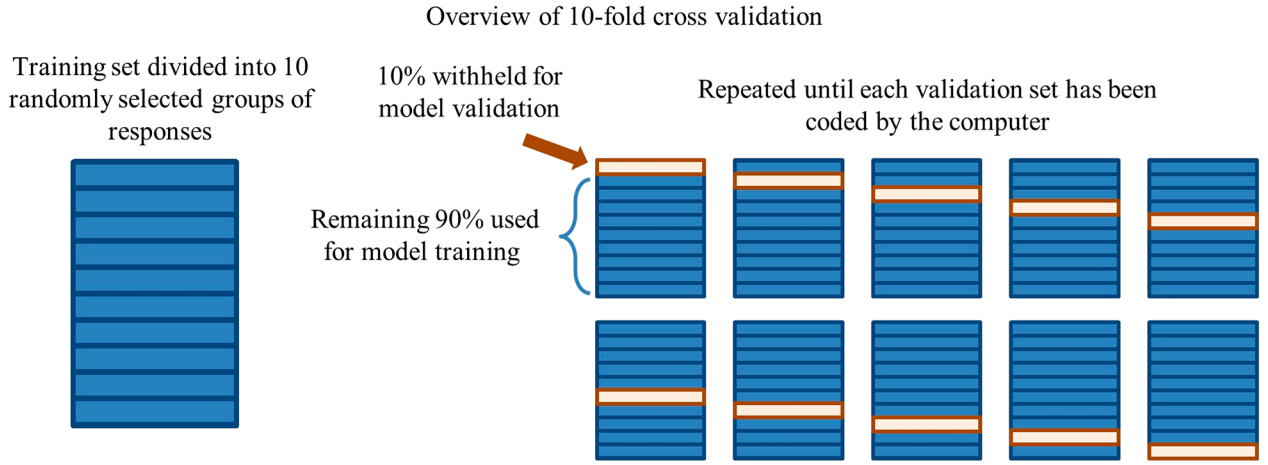

Figure 4. Overview of 10-fold cross-validation procedure for assessing the accuracy of the developed computer model.

matrix to classify each of the new responses into one of our three codes. Whichever code is picked by the most models is the machine's final consensus code for the response. In the unlikely event of a tie, the code assigned to the response is the first defined code with the maximum number of votes. For this coding scheme, the codes were defined in the following order: NE, EC, CM.

One way we assessed the accuracy of the computer model is through a 10-fold cross-validation (Figure 4). This cross-validation method provides insight into the accuracy of the model without the need for any additional human coded responses; that is, only the human coded training set is needed. In the cross-validation (which is automatically carried out by the CRC), the training set is divided into 10 randomly selected subgroups, each corresponding to 10% of the training set. Each subgroup is then coded by a new computer model trained with the remaining 90% of responses. This ensures that the same responses are not used simultaneously in the computer model training and testing. This process is then repeated a total of 10 times until the entire training set has received a computer predicted code. The CRC then calculates the agreement (using several statistics like Cohen's kappa) between the human and computer predicted codes generated in the cross-validation. By using the cross-validation built into the CRC, we can get an idea about how accurate the final computer model (trained on *all* the responses in the training set) would be without human coding any additional data. In this study, we also conducted additional tests to ensure that the final computer model is accurate. For this additional testing, we used the final computer model to predict codes for new sets of human coded responses from groups 1B, 2, and 3 which were not included in the training set and calculated the agreement between human and computer coding.

RESULTS AND DISCUSSION

Developing an Initial Model to Characterize LDF Responses

The first stage in the development of a robust model was to use 150 group 1A responses coded as part of a previous study[10] to train the computer model. Before training the computer model, we used the spellcheck feature in Microsoft Excel to identify and fix errors in the responses (correcting misspelled words, deleting duplicate words) to help the computer evaluate the words present in the responses rather than the spelling. The accuracy of this first model compared to the human coding using the cross-validation procedure was "moderate", reaching a Cohen's kappa of 0.58.[42] This initial stage did not have enough responses to develop an accurate model. More responses were needed to increase the lexical diversity of the training set to help the model better identify the patterns of words associated with each code. Author K.N. coded additional sets of 100 group 1A responses (randomly selected using a random number generator) to include in the training set. After the addition of 100 more responses, we trained a new computer model and found that the Cohen's kappa value (calculated from the cross-validation) had increased to 0.63. Each set of 100 responses was added iteratively to the training set, to train new models and assess the new model's performance using the cross-validation procedure (Figure 5).

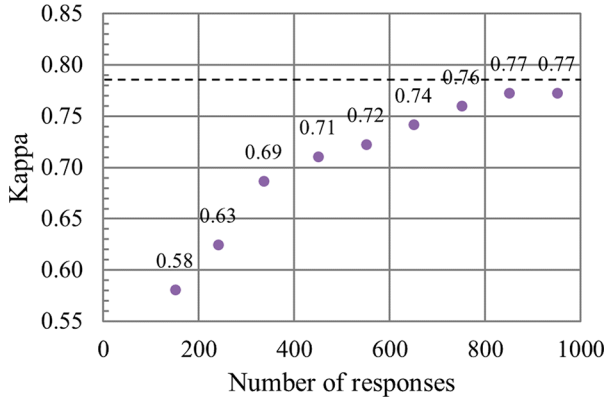

Figure 5. Agreement between the human and computer described by the Cohen's kappa value calculated in the 10-fold cross-validation as a function of the size of the number of responses used to train (and also validate) the computer model. The dashed line at 0.78 indicates the Cohen's kappa value for the human−human IRR with the responses.

This process continued until no more meaningful improvement of the cross-validation Cohen's kappa value was observed. After 950 total responses were coded, the Cohen's kappa from cross-validation reached 0.77, signifying "substantial agreement" between the computer and human coding.[42] Recall that the initial human IRR was 0.78 for group 1A responses. Williamson et al. propose that one metric for a successful model is having a quadratic weighted kappa

https://dx.doi.org/10.1021/acs.jchemed.0c00445
J. Chem. Educ. 2020, 97, 3923−3936

greater than 0.7.[53] In this study, we treated these categories as nominal rather than ordinal consistent with our assumptions from our previous study;[10] therefore, we did not use quadratic weighted kappa for our data. Our Cohen's kappa measure, a more conservative estimate, was above this target for model performance.

In this paper, we will refer to the computer model trained on the 950 group 1A responses as the "initial model". A crosstab illustrating the coding agreement between the human and computer codes (determined in the cross-validation) for the initial model is shown in Table 3. Of the 950 coded, 813 were coded the same by both the human and computer, corresponding to a proportion of 0.86 (accuracy value). Although the human and computer disagreed on 137 responses, these disagreements occurred in a mostly symmetric manner. For example, while the computer coded 35 EC responses as NE, it also coded 30 NE responses as EC. The result was that these disagreements had a smaller impact when considering the overall distribution of responses for the entire group.

Table 3. Crosstab Relating the Number (and Percentage of Total) of Reference Human Scores to the Predicted Computer Scores for the Training Set of the Initial Model

Initial model training set		Human reference scores			
		NE	EC	CM	Sum
Computer predicted scores	NE	178 (18.7%)	35 (3.7%)	4 (0.4%)	217
	EC	30 (3.2%)	379 (39.9%)	46 (4.8%)	455
	CM	1 (0.1%)	21 (2.2%)	256 (26.9%)	278
	Sum	209	435	306	950

Besides Cohen's kappa and accuracy, the sensitivity and specificity values provided important information about the performance of each of the bins in the model. Sensitivity is the proportion of correctly scored positive cases by the computer model, and specificity is the proportion of negative cases correctly scored by the computer model. For example, of the 209 human coded nonelectrostatic responses, the computer correctly coded 178 of those responses corresponding to a proportion of 0.85 (sensitivity value). Additionally, of the 741 responses that the human did not code as NE (i.e., coded as EC or CM instead), the computer coded 702 of those responses as not NE resulting in a specificity value of 0.95. For this model, all bins had both a high sensitivity (ranging from 0.84 to 0.87) and high specificity (ranging from 0.85 to 0.97). All of these factors indicated that the model was sufficiently trained and ready to characterize new responses, so long as they were also from this group of students. To make sure that our model had captured all the ways a student might explain this LDF causal mechanistically, we needed to include responses in our training set from other groups of students.

Expanding Our Initial Model with Responses from 3 Other Groups

Now that we had a model that was working well with responses in a single context, we needed to know if the model performed well with different groups of students who may have approached this task differently or had different vocabularies. Using the responses from groups 1B, 2, and 3, we explored how the model performed with groups of students who differed from each other in terms of when they were taking general chemistry, their general chemistry curriculum, the circumstances under which they responded to the task, and the racial/ethnic makeup of the group.

After authors R.L.M. and M.N. coded the responses from these three groups, we set aside 100 randomly selected responses from each group for later testing of the machine learning models. The rest of the responses were added to the initial model to create a new computer model (N = 1,730) which we call the "combined model" in this paper. As before, we spellchecked the responses included in the training set to give the algorithms the best chance of identifying the important patterns relevant to each category. With responses from a variety of different groups in the training set, the resulting combined model may be better able to capture other ways students explain this phenomenon that were previously not captured with the initial model.

On the basis of the cross-validation, the agreement between the combined model and the human scoring was very similar to that of the initial model (Table 4); the Cohen's kappa and accuracy values (0.78 and 0.86 respectively) did not change much, and the ranges of sensitivity and specificity values (sensitivity, 0.82−0.89; specificity, 0.86−0.97) were very similar to those of the initial model. On the basis of these metrics, it seemed that the combined model performed well but no better than the initial model. This might be because more than half of the responses in this training set came from a single group of students (1A). This meant that metrics evaluating model performance from the cross-validation were primarily reflective of how the model codes responses from group 1A, making it harder to see how the new combined model was able to predict responses from groups 1B, 2, and 3. To get a better sense of how the combined model fared compared to the initial model for those groups, we used both models to score the sets of responses from groups 1B, 2, and 3 withheld from the combined model.

Table 4. Crosstab Relating the Number (and Percentage of Total) of Reference Human Scores to the Predicted Computer Scores for the Training Set of the Combined Model

Combined model training set		Human reference scores			
		NE	EC	CM	Sum
Computer predicted scores	NE	513 (29.7%)	67 (3.9%)	5 (0.3%)	585
	EC	64 (3.7%)	633 (36.6%)	71 (4.1%)	768
	CM	1 (0.06%)	39 (2.3%)	337 (19.5%)	377
	Sum	578	739	413	1730

Testing the Model Performance

To conduct this test, we used the 100 coded responses from each group (1B, 2, and 3) that had already been human coded

https://dx.doi.org/10.1021/acs.jchemed.0c00445
J. Chem. Educ. 2020, 97, 3923−3936

but not included in the training set. The responses used to test the models were not spellchecked to simulate how an instructor might apply this tool in practice, where raw students' responses may be used. We report the agreement between the computer model predictions and the human consensus scores in Table 5. Note that the computer scores are compared to the human consensus scores; in other words, the final codes that authors R.L.M. and M.N. agreed upon after discussion. Their initial agreement (before discussing the responses) for each test set is also reported in Table 5. We included additional information about the alignment of human codes with both the initial and combined models in Supporting Information Section S3.

Table 5. Cohen's Kappa Value and Percent Agreement between Human Coders and with the Computer Models for Groups 1B, 2, and 3

Group	Human−Human Agreement	Human Consensus−Initial Model Agreement	Human Consensus−Combined Model Agreement
1B	0.74	0.74	0.74
N = 100	(83%)	(83%)	(83%)
2	0.72	0.67	0.72
N = 100	(89%)	(86%)	(89%)
3	0.64	0.77	0.80
N = 100	(78%)	(86%)	(88%)

For group 1B, both the combined and initial models showed the same level of agreement with the human consensus scores. This made sense considering that the initial model was trained with group 1A responses, which were from students attending the same institution and taking the same general chemistry curriculum; they differ primarily in the semester that they took the course. For group 2, the combined model had slightly better agreement with human consensus scores than the initial model. The combined model correctly scored an additional three responses, all within the nonelectrostatic category. The majority of group 2 responses were classified as non-electrostatic, so it could be that the addition of the group 2 responses in the combined model better allowed this model to correctly characterize nonelectrostatic responses.

For group 3, the combined model again showed higher level of agreement with the human consensus scores compared to the initial model, raising the Cohen's kappa value to 0.80. Interestingly, even for the initial model, the Cohen's kappa value (0.77) was quite a bit larger than the value for the agreement between the human coders for this same set (0.64). The lower level of agreement between the human coders was due to disagreements about how to code vague responses that were "edge cases" between the nonelectrostatic and electrostatic causal bins. It is promising that both computer models handled these new differences between the NE and EC bins well and that the addition of the group 3 responses to the combined model continued to improve its performance.

Further investigation of the responses misclassified by the combined model revealed that the computer struggled with the same responses as the human coders. If we consider the 300 test set responses from groups 1B, 2, and 3, the combined model misclassified 40 responses compared to the human consensus codes. Of those 40 responses, the human coders initially disagreed on how to code 19 of those responses (47.5%). Meanwhile, if we consider the remaining 260 responses to be correctly classified by the combined model, the human coders initially disagreed on only 31 of those responses (11.9%), a much lower proportion. Looking further into the responses misclassified by the combined model, we did not find that they were occurring primarily with any one particular code (see Supporting Information Section S3). We also examined the content of those misclassified responses and found that the bulk were either "edge cases" or a particularly atypical explanation.

Notably, all the computer models tested showed little to no degradation in agreement to human scores when compared to human−human agreement. In other words, the Cohen's kappa value for the human−computer model agreement was not much different than the Cohen's kappa value for the human−human agreement. This was particularly true for the combined model which performed very near or above the human−human agreement level. Even the model showing the largest degradation from human−human agreement measures, the initial model for group 2 students, still performed at an acceptable level, as the degradation was less than a suggested threshold of 0.1 difference (see Williamson et al.[53]).

All of these results suggest that the combined model, with additional responses from groups 1B, 2, and 3, can characterize student responses like human coders would, even improving upon the accuracy of the initial model coding of groups 2 and 3. While these improvements are modest, it does seem that they are the result of coding the responses from these different groups of students who may explain this phenomenon in ways we had not captured before. This is supported by the fact that the gains in agreement were only seen with groups 2 and 3, whose responses likely differ the most from group 1A. The high level of agreement between the computer and human coding for both of these groups is noteworthy. The students in group 2 are using a different curriculum in which they are not explicitly taught to generate causal mechanistic explanations, but still the computer model works. The high level of accuracy of the computer codes for group 3 is also important because this group of students is primarily Hispanic while the bulk of the other students in our computer model training and testing are White. This aligns with the joint recommendations for educational testing put forth by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, that in developing methods of scoring constructed responses (in particular automated scoring methods) we must be cognizant of the different subgroups of students in our testing populations and work to ensure that these resources are valid for all subgroups.[54] This is some evidence that these resources provide an equitable approach to the automated analysis of explanations; that is, these resources provide meaningful and accurate information that aligns with the human coding for diverse student populations.

Detecting Signal at the Group Level

While we have achieved good agreement between the humans and the combined model, we have not reached perfect agreement (although we note that the agreement is at least as good as human−human coding). Although there are bound to be errors in codes of individual student responses, the overall distribution of codes in a group (e.g., a class) may still be informative if the model is accurate overall and errors in misclassifications are symmetrical (see Tables 3 and 4 and Figure 6). In such a case, the predicted distribution of the group may only be minimally affected by errors and may still

https://dx.doi.org/10.1021/acs.jchemed.0c00445
J. Chem. Educ. 2020, 97, 3923−3936

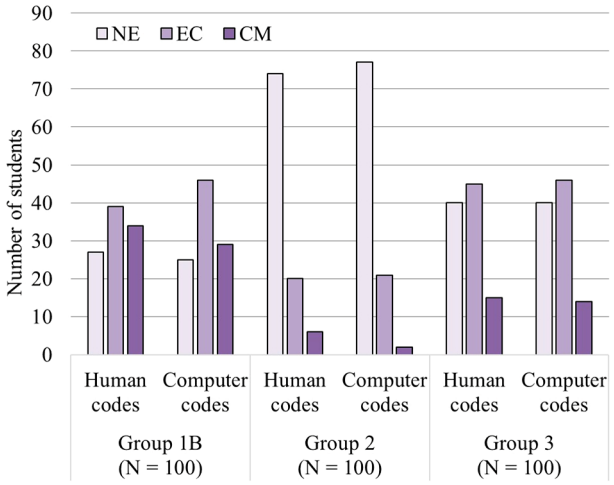

Figure 6. Distribution for LDF codes for the subsets of 100 responses from groups 1B, 2, and 3 as coded by humans (consensus score) and the computer (combined model). The human consensus score represents the agreed upon codes of authors R.L.M. and M.N. after discussion. The computer scores were coded by the combined model which was developed using responses from groups 1A, 1B, 2, and 3.

be valuable for instructors, even if some specific individuals have been misclassified. That is why in Figure 6 the human and computer codes look nearly identical despite there being disagreements on individual responses (see Supporting Information Section S3 for statistical analyses). For this reason, these resources should only be used for giving the instructor group-level information and should not used for high-stakes individual student information (i.e., points for providing a causal mechanistic response on a summative assessment).

Characterization of how a group of students responds to the prompt can give us important information about how that group is able to explain this phenomenon. Instructors can then use this information to better understand what their students can do and respond accordingly to better support their learning.[3] Such models should be able to detect different patterns of responses, for example, from different groups of students. Indeed, when we look at the group-level responses for the three test sets (Figure 6), we do see different distributions of responses for the different groups. For example, in a comparison of groups 1B and 2, there is a markedly different distribution of responses: there are more EC and CM responses in group 1B compared to group 2, which is mainly NE responses. We cannot say the exact cause for this difference as the prompt administration differed between the two groups. What we can say though is that, regardless of the cause, our model can detect that not many group 2 students are providing EC or CM responses. For the group 3 students, we see more CM and EC responses than are present for group 2. Again, we cannot say the cause of the difference, but our model is able to capture that there is a difference.

It is also worth reiterating that the combined model coding looks almost identical to the human coding. Not only is the model performing accurately, but also by coding responses from these other groups we have ensured that (1) their responses are part of the model we have developed and (2) our final combined model has been tested with responses from these different groups. That is, the resources we have developed are working well for a greater diversity of students in a variety of contexts compared to our initial model.

LIMITATIONS

As technology continues to develop and grow more sophisticated, the ability to automate the analysis of student assessment will improve. It is important to remember, however, that these tools cannot replace the human role in analysis outright. While computers can identify patterns in responses, we need expert input to determine if these patterns are meaningful. The importance of collaboration in developing reliable and efficient computer resources that are meaningful to the chemistry education community cannot be understated. This is a growing field where both current advances in the education and machine learning domains are taken into account. Additionally, we must understand that this technology is not perfect. We agree with other automated analysis experts that we should be hesitant about using such resources for making high-stakes decisions, in particular at the level of the individual response.[23,24] Instead, we view the most appropriate use of these technologies for providing feedback to instructors to better support student learning.

We acknowledge that the prompt asked students to provide both a drawing and text response, meaning that students' text explanations are only one part of the story. It is possible that students provided additional thinking in their drawings that was not captured in their text responses. We hope that in the future there could be a corresponding analysis of student drawings. For now, we hope that instructors can use a survey of student drawings along with the results of the automated analysis of students' text responses to inform their teaching.

Unfortunately, because we know that changes in prompt structure change the response it elicits,[20] this computer model should be limited to use with this specific prompt. It remains to be seen if these developed scoring models could still work with different interacting neutral entities (like argon). However, machine learning resources that automate the analysis of prompts exploring other phenomena can certainly be developed using the CRC. On the basis of our experience, developing these resources requires a large number of student responses, prior studies on prompt development and associated coding, and a great deal of time spent on human coding. This may limit the creation of more machine learning models, particularly by busy faculty in charge of teaching these courses. These faculty can still use other models and questions developed by other researchers on the AACR website, but it may serve as a barrier for the addition of more models.

Coding and adding more models from different populations of students improved the model performance for responses from groups 2 and 3. It may be that coding more responses from other populations of students would increase the model performance further. At some point, however, it will be necessary to stop adding more responses, when the time and effort it takes to collect and code more responses does not justify marginal gains in model performance (see Figure 5). However, this does not eliminate the need to validate coding carried out by this model, in particular with new populations of students that have not yet been included in the model development.

One further limitation is that such analyses are not currently able to provide the kind of individualized formative feedback that a human reader might (if they had the time). As noted earlier, in our large enrollment classes, feedback from these

https://dx.doi.org/10.1021/acs.jchemed.0c00445
J. Chem. Educ. 2020, 97, 3923−3936

analyses is provided to the group, rather than individually, and students are encouraged to reflect on and rework their responses. While this is not a substitute for the kind of Socratic dialogue that might be ideal, at the present time we do not have the capability for this kind of interaction.

CONCLUSION AND IMPLICATIONS FOR TEACHING AND RESEARCH

By using machine learning, we have developed a tool that instructors can use to better understand how their students can explain the origins of LDFs, an intermolecular force that is important throughout chemistry and biology. With knowledge of what their students can do, instructors can then modify their teaching practices to better support their students' learning. Additionally, with the ability to process large amounts of data, departments could use this tool to understand the impact of the instruction across different courses or time.

The use of open-ended explanatory questions is one of the few instructional techniques that is supported by strong evidence[16] (that is, multiple studies across multiple populations), yet for some institutions, especially in lower-level or high-enrollment courses, this approach may be ruled out because of the huge commitment of time and personnel that are needed to grade or evaluate such tasks. Here we show how analysis of one such task can be automated, and by including a range of institutions, student demographics, and curricula we have developed a model that appears to be more robust than simply using student data from one cohort. Additionally, the coding scheme we automated is not just picking out predefined keywords, but also it is capturing the patterns in the text response that correspond to different types of student reasoning. That is, we are characterizing more than just the presence of a student idea, but also how they use that idea as well. This speaks to the power of the CRC's ability to mimic sophisticated human coding based on the lexical patterns in the students' responses.

While this item and associated scoring model are now available for use (beyondmultiplechoice.org), it should be noted that a great deal of time and resources were expended not only in the human coding of the training data that made the model so robust, but also in the design of the prompt that elicited student reasoning. Clearly it is not feasible for individual instructors to design their own assessments and expect such results without similar expenditures. However, if researchers collaborate and pool their data for various tasks, it is feasible to build up a library of items and associated models where instructors can input their own student data.

The increased availability of these kinds of models means that, even for large enrollment courses, assessments need not be limited to forced choice, calculations, or one-word answers that tend to emphasize fragmentary or rote knowledge. The very act of constructing deep explanatory responses is linked to the development of more robust knowledge frameworks, and the more often students are asked to engage in this kind of activity, the more useful their knowledge will become. Answering these kinds of questions, where reasoning is necessary, requires that students understand the material. As has been shown in numerous research studies, the use of vocabulary terms alone does not necessarily correspond with understanding; it is instead the constructed reasoning responses that tend to elicit evidence of understanding. More opportunities to engage in this kind of activity can only improve learning.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.0c00445.

Additional information about the participant demographics, example student responses, and human and computer coding of groups 1B, 2, and 3 test sets (PDF, DOCX)

AUTHOR INFORMATION

Corresponding Author

Keenan Noyes − *Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0002-8587-1694; Email: noyeskee@msu.edu

Authors

Robert L. McKay − *Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0003-1925-5655

Matthew Neumann − *Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, United States*

Kevin C. Haudek − *Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0003-1422-6038

Melanie M. Cooper − *Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0002-7050-8649

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jchemed.0c00445

Notes

Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to thank the researchers and instructors at each institution for their time and effort in helping us to administer these activities. Additionally, this work would not have been possible without all the students who allowed us to use their responses for this work; we greatly appreciate their generosity. This work is supported by the National Science Foundation under DUE 1323162, DUE 1359818, DUE 1420005, and DUE 1341987.

REFERENCES

(1) Mislevy, R. J.; Almond, R. G.; Lukas, J. F. *A Brief Introduction to Evidence-Centered Design*; Research Report RR-03-16; Educational Testing Service, 2003.

(2) *Knowing What Students Know: The Science and Design of Educational Assessment*; National Academies Press: Washington, DC, 2001; p 10019. DOI: 10.17226/10019.

(3) Shepard, L. A. The Role of Assessment in a Learning Culture. *Educational researcher* **2000**, *29* (7), 4−14.

(4) Maeyer, J.; Talanquer, V. The Role of Intuitive Heuristics in Students' Thinking: Ranking Chemical Substances. *Sci. Educ.* **2010**, *94* (6), 963−984.

(5) 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.
(6) Cooper, M. M.; Williams, L. C.; Underwood, S. M. Student Understanding of Intermolecular Forces: A Multimodal Study. *J. Chem. Educ.* **2015**, *92* (8), 1288−1298.
(7) Barker, V.; Millar, R. Students' Reasoning about Basic Chemical Thermodynamics and Chemical Bonding: What Changes Occur during a Context-Based Post-16 Chemistry Course? *International Journal of Science Education* **2000**, *22* (11), 1171−1200.
(8) Taber, K. S. Building the Structural Concepts of Chemistry: Some Considerations from Educational Research. *Chem. Educ. Res. Pract.* **2001**, *2* (2), 123−158.
(9) Becker, N.; Noyes, K.; Cooper, M. M. Characterizing Students' Mechanistic Reasoning about London Dispersion Forces. *J. Chem. Educ.* **2016**, *93* (10), 1713−1724.
(10) Noyes, K.; Cooper, M. M. Investigating Student Understanding of London Dispersion Forces: A Longitudinal Study. *J. Chem. Educ.* **2019**, *96* (9), 1821−1832.
(11) National Research Council. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*; National Academies Press: Washington, DC, 2012. DOI: 10.17226/13165.
(12) Krist, C.; Schwarz, C. V.; Reiser, B. J. Identifying Essential Epistemic Heuristics for Guiding Mechanistic Reasoning in Science Learning. *Journal of the Learning Sciences* **2019**, *28* (2), 160−205.
(13) Nehm, R. H.; Schonfeld, I. S. Measuring Knowledge of Natural Selection: A Comparison of the CINS, an Open-Response Instrument, and an Oral Interview. *J. Res. Sci. Teach.* **2008**, *45* (10), 1131−1160.
(14) Hubbard, J. K.; Potts, M. A.; Couch, B. A. How Question Types Reveal Student Thinking: An Experimental Comparison of Multiple-True-False and Free-Response Formats. *LSE* **2017**, *16* (2), No. ar26.
(15) Lee, H.-S.; Liu, O. L.; Linn, M. C. Validating Measurement of Knowledge Integration in Science Using Multiple-Choice and Explanation Items. *Applied Measurement in Education* **2011**, *24* (2), 115−136.
(16) Pashler, H.; Bain, P. M.; Bottge, B. A.; Graesser, A.; Koedinger, K.; McDaniel, M.; Metcalfe, J. Organizing Instruction and Study to Improve Student Learning. *IES Practice Guide. NCER 2007-2004*; US Department of Education, 2007. DOI: 10.1037/e607972011-001.
(17) Bishop, B. A.; Anderson, C. W. Student Conceptions of Natural Selection and Its Role in Evolution. *J. Res. Sci. Teach.* **1990**, *27* (5), 415−427.
(18) Crandell, O. M.; Kouyoumdjian, H.; Underwood, S. M.; Cooper, M. M. Reasoning about Reactions in Organic Chemistry: Starting It in General Chemistry. *J. Chem. Educ.* **2019**, *96* (2), 213−226.
(19) Crandell, O. M.; Lockhart, M. A.; Cooper, M. M. Arrows on the Page Are Not a Good Gauge: Evidence for the Importance of Causal Mechanistic Explanations about Nucleophilic Substitution in Organic Chemistry. *J. Chem. Educ.* **2020**, *97* (2), 313−327.
(20) Cooper, M. M.; Kouyoumdjian, H.; Underwood, S. M. Investigating Students' Reasoning about Acid−Base Reactions. *J. Chem. Educ.* **2016**, *93* (10), 1703−1712.
(21) Black, P.; Wiliam, D. Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice* **1998**, *5* (1), 7−74.
(22) Sadler, D. R. Formative Assessment: Revisiting the Territory. *Assessment in Education: Principles, Policy & Practice* **1998**, *5* (1), 77−84.
(23) Dood, A. J.; Fields, K. B.; Raker, J. R. Using Lexical Analysis To Predict Lewis Acid−Base Model Use in Responses to an Acid−Base Proton-Transfer Reaction. *J. Chem. Educ.* **2018**, *95* (8), 1267−1275.
(24) Dood, A. J.; Dood, J. C.; Cruz-Ramírez de Arellano, D.; Fields, K. B.; Raker, J. R. Analyzing Explanations of Substitution Reactions Using Lexical Analysis and Logistic Regression Techniques. *Chem. Educ. Res. Pract.* **2020**, *21* (1), 267−286.
(25) Haudek, K. C.; Prevost, L. B.; Moscarella, R. A.; Merrill, J.; Urban-Lurain, M. What Are They Thinking? Automated Analysis of Student Writing about Acid-Base Chemistry in Introductory Biology. *Cell Biology Education* **2012**, *11* (3), 283−293.
(26) Kaplan, J. J.; Haudek, K. C.; Ha, M. Using Lexical Analysis Software to Assess Student Writing in Statistics. *Technology Innovations in Statistics Education* **2014**, *8* (1), https://escholarship.org/uc/item/57r90703.
(27) Prevost, L. B.; Smith, M. K.; Knight, J. K. Using Student Writing and Lexical Analysis to Reveal Student Thinking about the Role of Stop Codons in the Central Dogma. *LSE* **2016**, *15* (4), No. ar65.
(28) Sripathi, K. N.; Moscarella, R. A.; Yoho, R.; You, H. S.; Urban-Lurain, M.; Merrill, J.; Haudek, K. Mixed Student Ideas about Mechanisms of Human Weight Loss. *LSE* **2019**, *18* (3), No. ar37.
(29) Automated Analysis of Constructed Response. https://beyondmultiplechoice.org/ (accessed Mar 26, 2020).
(30) Haudek, K. C.; Kaplan, J. J.; Knight, J.; Long, T.; Merrill, J.; Munn, A.; Nehm, R.; Smith, M.; Urban-Lurain, M. Harnessing Technology to Improve Formative Assessment of Student Conceptions in STEM: Forging a National Network. *Cell Biology Education* **2011**, *10* (2), 149−155.
(31) Sieke, S. A.; McIntosh, B. B.; Steele, M. M.; Knight, J. K. Characterizing Students' Ideas about the Effects of a Mutation in a Noncoding Region of DNA. *LSE* **2019**, *18* (2), No. ar18.
(32) Jurka, T. P.; Collingwood, L.; Boydstun, A. E.; Grossman, E.; Atteveldt, W. V. *RTextTools: Automatic Text Classification via Supervised Learning*; 2012.
(33) Nehm, R. H.; Ha, M.; Mayfield, E. Transforming Biology Assessment with Machine Learning: Automated Scoring of Written Evolutionary Explanations. *J. Sci. Educ. Technol.* **2012**, *21* (1), 183−196.
(34) Opitz, D.; Maclin, R. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research* **1999**, *11*, 169−198.
(35) Ha, M.; Nehm, R. H.; Urban-Lurain, M.; Merrill, J. E. Applying Computerized-Scoring Models of Written Biological Explanations across Courses and Colleges: Prospects and Limitations. *LSE* **2011**, *10* (4), 379−393.
(36) Cooper, M. 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.
(37) Brown, T. L.; LeMay, E. H., Jr.; Bursten, E. B. E.; Murphy, C. J.; Woodward, P. M.; Stoltzfus, M. W. *Chemistry: The Central Science*, 14th ed.; Pearson: New York, 2018.
(38) Lage, M. J.; Platt, G. J.; Treglia, M. Inverting the Classroom: A Gateway to Creating an Inclusive Learning Environment. *Journal of Economic Education* **2000**, *31* (1), 30−43.
(39) Bryfczynski, S. *BeSocratic: An Intelligent Tutoring System for the Recognition, Evaluation, and Analysis of Free-Form Student Input*. Ph.D. Dissertation, Clemson University, Clemson, SC, 2012.
(40) *Qualtrics*; Qualtrics: Provo, UT.
(41) Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20* (1), 37−46.
(42) Landis, J. R.; Koch, G. G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33* (1), 159.
(43) Jurka, T. P.; Collingwood, L.; Boydstun, A. E.; Grossman, E.; van Atteveldt, W. RTextTools: A Supervised Learning Package for Text Classification. *R Journal* **2013**, *5* (1), 6−12.
(44) Feinerer, I.; Hornik, K.; Meyer, D. Text Mining Infrastructure in *R*. *Journal of Statistical Software* 2008, *25* (5). DOI: 10.18637/jss.v025.i05.
(45) Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. Support Vector Machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13* (4), 18−28.
(46) Mcauliffe, J. D.; Blei, D. M. Supervised Topic Models. *Advances in Neural Information Processing Systems* **2008**, 121−128.
(47) Friedman, J. H.; Hastie, T.; Tibshirani, R. Additive Logistic Regression: A Statistical View of Boosting. *Annals of Statistics* **2000**, *28* (2), 337−407.

https://dx.doi.org/10.1021/acs.jchemed.0c00445
J. Chem. Educ. 2020, 97, 3923−3936

(48) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, 1984.
(49) Hothorn, T.; Lausen, B. Bundling Classifiers by Bagging Trees. *Computational Statistics & Data Analysis* **2005**, *49* (4), 1068−1078.
(50) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.
(51) Friedman, J.; Hastie, T.; Tibshirani, R. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics* **2008**, *9* (3), 432−441.
(52) Kazama, J.; Tsujii, J. Evaluation and Extension of Maximum Entropy Models with Inequality Constraints. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics, 2003; Vol. *10*, pp 137−144. DOI: 10.3115/1119355.1119373.
(53) Williamson, D. M.; Xi, X.; Breyer, F. J. A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice* **2012**, *31* (1), 2−13.
(54) American Educational Research Association; American Psychological Association; National Council on Measurement in Education. Fairness in Testing. In *Standards for Educational and Psychological Testing*; 2014; pp 49−72.

https://dx.doi.org/10.1021/acs.jchemed.0c00445
J. Chem. Educ. 2020, 97, 3923−3936