Automated analysis of short responses in an interactive synthetic tutoring system for introductory physics

Christopher M. Nakamura,[1,*] Sytil K. Murphy,[1,†] Michael G. Christel,[2]
Scott M. Stevens,[2] and Dean A. Zollman[1]

[1]*Department of Physics, Kansas State University, Manhattan, Kansas 66506, USA*
[2]*Carnegie Mellon University, Entertainment Technology Center (ETC), Pittsburgh, Pennsylvania 15219, USA*

(Received 30 April 2014; revised manuscript received 11 October 2015; published 16 March 2016)

Computer-automated assessment of students' text responses to short-answer questions represents an important enabling technology for online learning environments. We have investigated the use of machine learning to train computer models capable of automatically classifying short-answer responses and assessed the results. Our investigations are part of a project to develop and test an interactive learning environment designed to help students learn introductory physics concepts. The system is designed around an interactive video tutoring interface. We have analyzed 9 with about 150 responses or less. We observe for 4 of the 9 automated assessment with interrater agreement of 70% or better with the human rater. This level of agreement may represent a baseline for practical utility in instruction and indicates that the method warrants further investigation for use in this type of application. Our results also suggest strategies that may be useful for writing activities and questions that are more appropriate for automated assessment. These strategies include building activities that have relatively few conceptually distinct ways of perceiving the physical behavior of relatively few physical objects. Further success in this direction may allow us to promote interactivity and better provide feedback in online learning systems. These capabilities could enable our system to function more like a real tutor.

DOI: 10.1103/PhysRevPhysEducRes.12.010122

I. INTRODUCTION

The Internet is continuing to emerge and evolve as an important enabling technology for education. It is clear that the Internet provides a fast, efficient means of distributing a range of multimedia content, allows people to discuss and propagate that content, and allows people to interact both with each other and with interactive environments embedded within websites. The Internet is already being used for online courses and distance learning, and these have been domains of considerable growth [1]. Accurately predicting continued growth is difficult, but there is considerable evidence that online instruction will continue to play an important role in education [1]. Machine learning and data mining have recently emerged as useful approaches for investigating learning processes [2,3]. Researchers have increasingly been applying these techniques to the development of computerized tutoring systems [4–6]. The development of computerized tutors, both intelligent and unintelligent, has been an active area of research for some time [7–10]. Research aimed at improving and broadening online learning experiences is clearly of current interest, and the work we present here has implications for providing interactive feedback in online learning environments of all types.

An area of particular interest recently has been Massive Open Online Courses (MOOCs), which allow very large numbers of people to enroll in the same course without being geographically near the institution offering the course [11,12]. There is evidence of significant attrition problems [11]. Interactive technologies like those explored in this paper may be useful for improving MOOCs.

Another important type of online educational technology that is already commonly used is the online homework system. These systems have emerged in many fields, including physics, as a means of allowing students maximum opportunity to develop and apply their knowledge and skills, while maintaining reasonable levels of grading responsibilities for instructors [13,14]. Automatic assessment of students' answers to numeric or multiple-choice questions is a critical component of these homework systems. Our work shows progress towards automated assessment of student responses to a wider range of

*Corresponding author.
cnakamur@svsu.edu
Also at Department of Physics, Saginaw Valley State University, University Center, Michigan 48710, USA.
†Also at Department of Environmental and Physical Sciences, Shepherd's University, Shepherdstown, West Virginia 25443, USA.

question types, which may ultimately be useful for online homework systems.

Tutoring via computer has also been a goal in physics education research [8,9]. Our work, which centers on the design, development, and testing of an online video-based synthetic tutoring system, has the potential to further extend the range of instructional possibilities offered online and do so in a way that aims to emulate one-on-one tutoring. Tutoring was selected as the mode of instruction for this project because of its well-known efficacy [15,16]. The system provides instruction via multiple multimedia-based components, but the centerpiece of the system is a video tutor built on synthetic interview technology [17]. This technology allows students to type natural language questions and receive prerecorded video responses to those questions [17]. It is clear, however, that a tutor does more than answer questions, and, as discussed below, our system provides more than just answers to questions. A key provision is feedback based on student performance. The results that we present here focus on understanding how to build a system that provides a more interactive and instructive experience by providing this feedback via automated assessment of students' text responses to short-answer questions.

To replicate the beneficial results of tutoring, it is instructive to understand the mechanics of tutoring and how they give rise to a positive instructional interaction. Considerable research has been conducted to understand the tutoring process and what makes it so effective [18–21]. While attributing the increase in efficacy to an increase in the teacher's ability to provide customized instruction is tempting, prior research suggests that the superior efficacy of tutoring as a teaching and learning method is due more to an enhanced opportunity for students to interact cognitively with the material to be learned [21]. A useful 5-step model of tutoring has been established [19]. We make use of that model in this work to identify roles that our synthetic tutor can or should play, and to help understand how it succeeds and fails in performing those roles. The five steps in that model are as follows:

(1) Tutor provides a problem or question.
(2) Student answers the question or solves the problem.
(3) Tutor provides targeted feedback on the answer or solution.
(4) Tutor and student collaborate to make further progress.
(5) Tutor assesses the student's knowledge.

While this model may not encompass every approach to tutoring, or the occurrences in every tutoring session, it does provide a useful framework for understanding many tutoring sessions, and provides a set of conditions a synthetic tutoring system should satisfy. Natural language processing technology and web-deployed video have enabled us to build a system that can present students with a problem or question and answer students' typed natural language questions regarding relevant physics content. That type of system can be effective for building a tutoring session that provides items 1, 2, and 4. Our system provides a video discussion of the activities after the student has completed them, which acts as a form of feedback, however, it is not focused and responsive to individual needs. It does not provide assessment of individual students' work. In that sense, we see that our system is less capable in providing items 3 and 5. This is a foreseeable limitation of our system, but it is not a limitation that is easily overcome without data. Once we understand how students respond to lesson questions we can begin to ask how the synthetic tutor should respond to the student in a way that provides useful feedback and allows for assessment of student learning. An important requirement for our system to function as a tutor is that the approach must be one that a computer can execute in real time. Providing students with feedback two days later after a person has read their work is not viable. In this paper we present the results of an analysis that used machine learning to classify students' responses to short answer questions based on the contents of their responses. The approach shows initial promise for allowing us to develop a more interactive synthetic tutor that could automatically identify the types of ideas being expressed by the student and respond in appropriate ways.

The benefits of automated assessing of student work are fairly obvious, and considerable effort has been directed towards that goal. Clearly the means of assessment differ depending on the type of work students submit. Computer systems can easily assess the numerical correctness of students'answers to questions or problems. Online homework systems in math and the sciences have hinged on this ease. Responses to multiple-choice questions are similarly easy. These types of responses, however, fail to convey significant components of student understanding, and automated assessment of longer, richer text responses is highly desirable. These text responses can be broken down, relatively neatly, into essays and short answers. Computerized essay scoring dates to the 1960s and continues to be an active area of research today [22–25]. Automated assessment of short text answers to questions is also an area of interest both in and out of science education [26–29]. Automatic assessment of short responses is in some sense more difficult because of the smaller quantity of text from which a program must make the assessment [25]. Some effort in this direction has focused on the classification of answers that can be objectively classified as right or wrong [27]. However, classification schemes that are more nuanced and reflect the similarities and differences in ideas expressed by students are clearly more interesting both to researchers who are interested in a more detailed picture of what students think about a given idea and to teachers who need that detailed picture to successfully intervene when students do not understand. Our research therefore focuses on classifying short-text responses to questions not based on correctness but based on common

ideas expressed in the responses. We believe this approach is an interesting and potentially productive one, worthy of investigation, for developing a system that works as a tutor. Furthermore, we believe that the approach may be useful to researchers and instructors in a broader sense. Automatic assessment has the potential to allow instructors to provide students with feedback in online learning environments, including MOOCs, that might look quite different from ours. There is already evidence that automated assessment of written work in the context of biology correlates reasonably well with assessment via oral clinical interviews [30]. There has long been evidence that mathematical assessment methods have distinct advantages worth pursuing [31]. It is logical to consider applying these techniques in our application, however, the approach must be investigated. If shown effective, it also has the potential to expedite grading. Therefore, while we investigate the techniques to advance our tutoring project, we stress their potential applicability far beyond that scope.

II. PATHWAY ACTIVE LEARNING ENVIRONMENT AND ITS FUNCTION

Our system, the Pathway Active Learning Environment (PALE), is a multimedia-based synthetic tutoring system designed to promote active learning of physics concepts with support from a natural-language video interface. The system is targeted at high school and college physics students studying algebra- or concept-based physics. The system's user interface is shown in Fig. 1. Our system provides prerecorded video responses to students' typed questions, without the use of artificial intelligence. This technology, termed the Synthetic Interview (SI), was developed as a means of simulating a conversation [17]. Students can enter natural language questions by typing

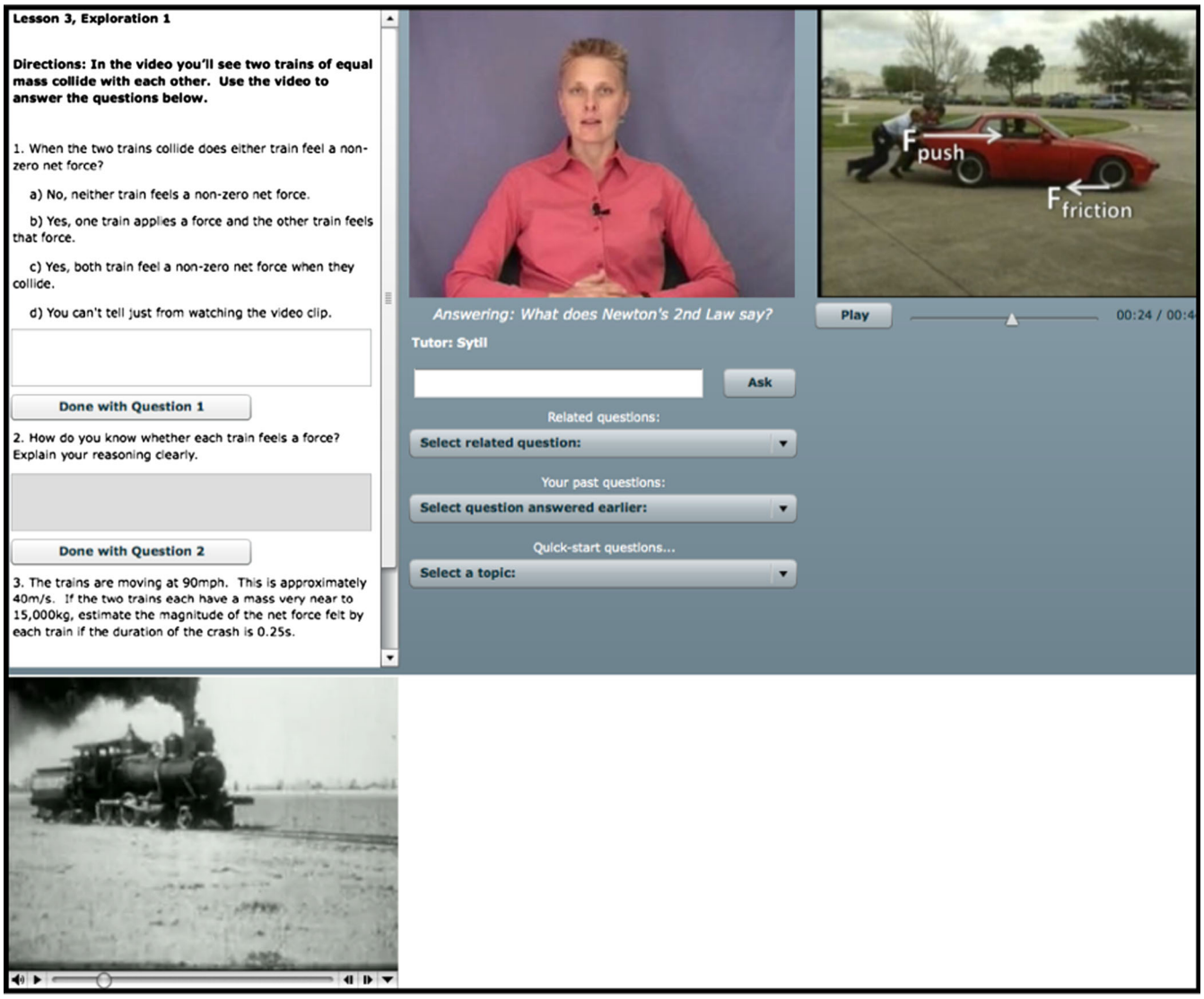

FIG. 1. The Pathway Active Learning Environment interface: The lesson activities and relevant video clip can be seen on the left. The SI tutor can be seen in the center of the interface. At the right a sample video supplementing the tutor's explanation is shown.

them into an input field. However, just as in a tutoring session, students may not always know what to ask. Thus, they are also provided with a set of so-called quickstart questions from which they can choose. The system interface is shown in Fig. 1. It can be broken up spatially into three panels: a left panel, a right panel, and a center panel. The video tutor component of the system, along with one of the SI tutors, is shown in the center panel. Questions are submitted beneath the video player, which provides the responses. The function of the components of the right and left panels are described in the next two paragraphs.

The model of tutoring described previously requires a question or problem for the student to work on. Our system provides this, in the form of lesson activities built around video clips. Video clips were chosen as the focus of the lesson activities because they provide direct, observable connections between the ideas of physics and the physical behavior of real objects. A related capability of the system is the possibility of extracting time information (by advancing the video clip frame by frame) and displacement information (via on-screen scales) directly from the video clip. This kind of application of video in physics education has a long, well-established record of use in physics education [32–36]. In this sense, direct extraction of quantitative information that precisely connects the ideas of physics to physical objects in the video is possible. Using information from the video, students who use the PALE system must answer both quantitative and conceptual questions about the video clip in each activity. The lesson questions and video clip serve as the context for the tutoring process and can be seen in the left panel of the interface in Fig. 1. The students read and answer the questions at the top of the screen. The video clip related to those questions is at the bottom of the screen on the left side.

Human tutors can use paper and pencil to positive effect in a tutoring session, drawing pictures, and sketching processes that occur in time. While it is not possible for our system to directly replicate this ability we can try to build a somewhat similar feature into the PALE system. We can do this by using additional multimedia to supplement the tutor's explanations of physics ideas. This facet of the system can be seen on the right side of the screen in Fig. 1. As the system currently exists, this multimedia can take the form of narrated video clips or static pictures. The goal of this media is to illustrate and clarify the SI tutor's responses to student questions.

To make use of established pedagogy we developed our lessons to follow a three-stage learning cycle [37]. In each lesson students begin with a series of three activities designed to help the students explore the ideas to be learned in the lesson. This was followed by a formal concept introduction stage that was provided via a video presentation from the tutor persona. Each lesson concluded with three activities designed to help students apply the ideas in new contexts. While it is not possible to *a priori* gauge students' knowledge and experiences prior to working with our system, we developed or selected video clips that were likely to connect to students' prior life experiences and prior knowledge. We have developed three lessons, one on each of Newton's laws.

Ultimately successful tutoring exercises should elicit student's current beliefs about subject matter, or their current approaches to solving problems in clear ways to allow the tutor to accurately gauge student understanding (as reflected in the 5-step model). To fulfill this role the majority of our video-based lesson activities require students to make a prediction (along with an explanation or justification of that prediction) about the behavior of one or more physical objects, then observe the behavior and explain it using the relevant ideas. This predict-observe-explain process is well established [38,39]. We believe that it represents a sound pedagogical structure that mirrors the methods of scientific inquiry, and also promotes the function of our multimedia system in a more tutorlike manner. It is related to the elicit-confront-resolve process [40].

To best study the utility and efficacy of the PALE system in helping students learn it is necessary to gain a detailed picture of how students use the system, which requires logging student interactions with the system along with an ID tag that allows us to tie each interaction to an individual. Therefore students who used the PALE system were required to create accounts and the system logged all queries to the tutor (along with information about how the query was submitted), along with their responses to the lesson questions. The PALE logs this information with a time stamp so it is possible to reconstruct a detailed picture of each student's interactions with the system. The responses analyzed for this work were extracted from this log.

III. RESEARCH DESIGN

In this paper we discuss research aimed at investigating the potential benefits of using machine learning for the automated assessment of students' short answers to conceptual physics questions in the context of our online learning environment. The central question that we seek evidence to answer is whether it is reasonable to train a computer to automatically classify a student's response to a short-answer question based on the contents of the response. To do this we need a relatively large corpus of responses to short-answer questions. It is well established that virtually all automated text analysis schemes work better with larger data sets [29,41–45]. However, the successful demonstration of an automated analysis scheme that can be performed on smaller data sets opens the technique to a much wider array of educators and researchers, not just those who can build large data sets. Here we investigate whether an automated analysis scheme can be applied to data sets that consist of around 100–150 responses. This represents a corpus size that virtually any instructor or researcher could obtain in a reasonable

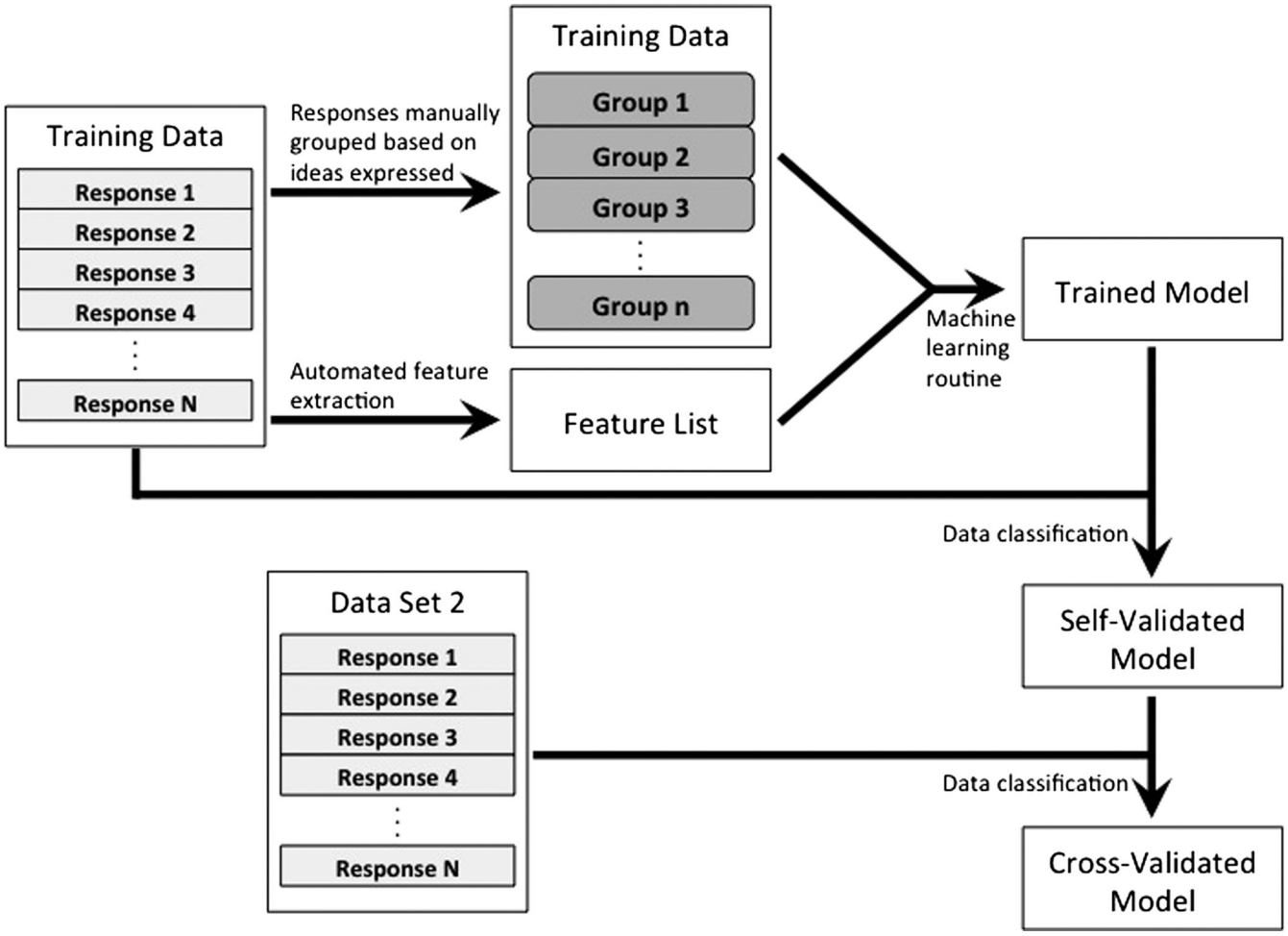

FIG. 2. Flowchart indicating the general analysis scheme followed in this paper.

time frame, and that larger-scale operations could obtain quite quickly and easily. We believe it also represents a lower limit on the size of a data set for which this type of analysis is feasible. We note prior research on using computerized scoring with short-answer questions [28,29]. These efforts focus, quite reasonably, on very large data corpuses because large corpuses generally produce more robust, effective computer models [29,41–45]. Since collecting large data sets requires effort, particularly for high school teachers and instructors at small colleges and universities, understanding how such techniques can be applied to smaller data sets is of interest, and has not been well explored.

To obtain the data for our analysis we collected responses from a variety of students who worked with the PALE system under various conditions. We collected responses from 22 algebra-based physics students who used the system in our interview facility, 30 algebra-based physics students who used the system as part of a homework assignment in a university computer-equipped classroom, 99 concept-based physics students who used the system (at a time and location of their choosing) as part of an assignment from their course instructor, and 41 high school physics students who used the system in their classroom as part of an in-class assignment from their high school teacher. While these student populations are quite different in some ways, our system is generally appropriate for students in introductory high school and college physics classes. Moreover we see some evidence for similarities in the ideas expressed by students across these different student groups. We believe that it is therefore appropriate to combine response sets from these different populations to better produce a system that can respond to a wider range of students.

A limited amount of demographic information was collected for the students via self-reporting prior to completing the exercises. We know, for example, that almost all of the students indicated that they self-identify as white and that the sample contains more females than males (the conceptual physics class targets elementary education majors, thus this makes sense). For our purposes this type of information was not of highest importance. However, in retrospect, there is demographic information that might be of some importance that was not collected, such as the number of non-native English speakers or English language

learners were present in the sample. Unfortunately, we cannot comment on this in precise, quantitative ways but given the demographics of Kansas State University and the high schools from which we sampled we have no reason to believe that a significant portion of the students are English language learners or non-native English speakers.

To conduct our analysis we used a utility designed for this kind of analysis called LightSIDE [46]. LightSIDE allows users to extract feature lists from text data, and, using user-provided coding schemes, train computer models that can be used to code more text data [46]. LightSIDE is, at the time of this writing, free for download, which makes it a low-cost utility for this type of analysis [47]. LightSIDE (or its predecessor, SIDE) has been used in prior research that exploits machine learning for computerized assessment of student work [29,30,48,49]. Its choice is justified as a robust, time-efficient tool, well suited to this type of analysis, in some cases outperforming commercial software [48,49].

In this research we use an established supervised learning approach to classification of data [41]. A schematic of the analysis approach is shown in Fig. 2. In this approach we take a set of responses to a given question and begin by grouping them based on commonalities in the ideas that are expressed in the responses. This requires reading over the responses multiple times to gain insight into the range of ideas expressed. Multiple groupings for a given response set are possible, and some may be more insightful than others. As will be discussed in greater detail below, decisions about how coarsely or finely to disaggregate the responses can ultimately make differences on the success of the computer assessment. Once the responses have been manually grouped they are randomized and divided into two equal-sized groups. A feature list is extracted for each group. While any searchable facet contained within the group of responses can be considered a feature, in this work we focused on the single words that made up the sets of responses; and, therefore, the feature list in this work is a list of all the words that show up in a given response set. It is possible to extract pairs, triplets, or larger groups of words as features in machine learning (bigrams, trigrams, and n-grams, respectively). We have observed in our data sets that extracting these additional features adds significantly to the analysis time while failing to improve the results. Once feature lists are extracted for each half of the response set a machine-learning algorithm is used to train a computer model capable of classifying additional responses using the same classification scheme. The algorithm uses the feature list, the human researchers' groupings, and the distribution of the features within the response set to identify the features that differentiate responses in one group from responses in another group. A number of machine learning algorithms are useful for this kind of analysis; NaïveBayes and Support Vector Machines (SVM) are two that are commonly used [42–45]. We experimented with the implementations of both algorithms that are packaged with LightSIDE. The results obtained here were obtained using SVM.

A commonly used metric for assessing interrater agreement is Cohen's kappa [50]. The primary benefit of using Cohen's kappa over the simpler percent agreement is that the kappa statistic takes into account the possibility of random agreement. The statistic is calculated via

$$\kappa = \frac{P_{\text{obs}} - P_{\text{rand}}}{1 - P_{\text{rand}}}, \tag{1}$$

where P_{obs} is the actual observed rate of agreement between the two raters and P_{rand} is the expected rate of agreement, assuming purely random agreement. Clearly a kappa value of 0 is consistent with an observed rate of agreement equal to the expected rate of random chance agreement. Standards for interpreting the kappa statistic are context dependent, and a good result in one situation may not be satisfactory in another. Landis and Koch introduced a rubric for interpreting kappa in 1977, which is now commonly used [51]. In order to provide some objective standard for comparison we make use of this rubric; it is summarized in Table I.

Ultimately benchmarks for our interrater agreement standards must come from understanding about minimum standards of functionality for our system, and how it can be made to function increasingly like a human tutor. This requires an understanding of how accurate a human tutor can be in assessing student meaning, and tying our results to that figure of merit. We address this explicitly in the discussion section.

Not all questions resulted in response sets amenable to this type of analysis, and we should not expect this would be the case. Instead, it is expected that most questions will not result in response sets where we can use this scheme. In the context of a lesson activity, however, it may be sufficient if even just one question can be automatically assessed. This could provide a sufficient glimpse into what the student might be thinking so that productive instructional feedback can be given. Over the course of our three lessons, and 18 lesson activities students answered nearly 80 questions, yielding over 10 000 responses. Since our scheme requires a human to read and code the questions, we determined it was not reasonable to try to analyze all of

TABLE I. Interrater agreement rubric from [51].

κ value	Agreement
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.61	Moderate
0.61–0.80	Substantial
0.81–1.00	Near perfect

the responses we had collected. We, instead, looked for examples of questions where the analysis seemed likely to be appropriate. The questions were chosen based on whether clear, conceptually distinct groupings of responses could be identified within the total response set. While we do not expect every question to produce this type of conceptual coherence it is a desirable quality which we seek out. To analyze the data the total data set was reviewed and evaluated in a coarse manner to identify questions that were likely to produce conceptually distinct groupings. Once questions of this type were identified, the responses to each question were carefully and repeatedly reviewed via a qualitative analysis approach to determine the range of ideas expressed in each response set and how the responses best grouped together. We do not and cannot argue that the groupings are unique, nor that there is one best grouping of the responses. Once a response set was completely coded, the analysis depicted in Fig. 2 was performed using LightSIDE. Prior to this analysis the data were run through the Microsoft Office spell-checking program to minimize the possibility of mismatch due to spelling errors. The process was conducted by a combination of the spell checker and visual inspection to ensure that typographical errors were adjusted to proper spelling without changing the meaning (as well as could reasonably be determined). In two instances the intended word was unclear and the misspelled word was not changed. The rate of spelling errors within sets of responses to a given question ranged from 0.06 errors per response to 0.25 errors per response, with 0.2 errors per response typifying the overall corpus of responses. By far the most common issue was the spelling of because as "becuase" with alternate spellings of acceleration and cylinder also recurring in the response sets. Overall spelling and typographical errors were a minor nuisance, not a significant problem. A small subset of the data consisting of confusing, irrelevant, or profane responses were removed as well. In future work in this direction automatic spell checking and noise rejection would be both feasible and desirable. The results we present come from analysis conducted on sets of responses to nine questions taken from nine of the activities. To allow the reader to understand more of the context in which the responses were collected we will describe briefly each of the nine activities and questions.

(1) Ball and track: In this activity students explored Newton's first law in the context of a ball rolling along a flat, low-friction track. We present the analysis of responses that resulted from asking students how the speed of the ball (the motion of which they had already observed) would change if the track's length was doubled.
(2) Car and coffee cup: In this Newton's first law activity, students observed the motion of a coffee cup "accidentally" left on the back of a car as the car drives off. We present the analysis of responses that resulted from asking students how the coffee cup's velocity in the horizontal direction changes as the car begins to drive.
(3) Crash test dummy: In this Newton's first law activity, students observe, and quantitatively characterize the motion of an unrestrained crash test dummy during a crash. We present the analysis of responses that resulted from asking students why the dummy's motion ultimately stops.
(4) Coin and cylinder: In this Newton's first law activity, students were posed the problem of obtaining a coin stuck lightly inside a graduated cylinder. We present the analysis of students' proposed methods of obtaining the coin.
(5) Beaker and coin: In this Newton's first law activity, students were presented with a coin sitting on a card, which in turn was sitting on a beaker. Students were asked to predict, observe, and explain the motion of the coin as the card was quickly pulled. We present the analysis of the predictions.
(6) Hammer and feather: In this Newton's second law activity, students were asked to predict the motion of a hammer and feather released from rest by an astronaut on the moon. We present the analysis of responses that resulted from asking students to predict what would happen when the two objects were simultaneously released, and, in particular, how the two objects' motions would compare.
(7) Train crash: In this Newton's third law activity, students observed and quantitatively investigated the motion of two trains of equal mass colliding at equal and opposite velocities. We present the analysis of responses that resulted from asking students to explain how they could tell whether the trains felt forces during the collision.
(8) Bowling ball and ice skater: In this Newton's third law activity, students observed and quantitatively investigated the motion of a bowling ball and ice skater when the ice skater throws the bowling ball. The low-friction environment provided by the ice enables the students to see clearly how forces are exerted on the skater and the ball. We present the results of analyzing a response set that resulted from asking students which object experiences more acceleration and why.
(9) Live and dead ball: In this Newton's third law activity, students were asked to compare the respective forces felt by a bouncing and nonbouncing ball upon impact with a table. We present the analysis of responses that resulted from asking students which material they would prefer having dropped on their own chest.

There is variance in the length of the responses students provide, and this is clearly an important characteristic of the response sets. If students provide one word responses then

this is not a productive line of research, but also the responses should not be long essays. This can be well characterized. For example, in the ball and track activity the response set had a mean length of 100 characters with a standard error of 5 characters. This corresponds to a typical response that is 1 or 2 complete sentences in length. The maximum length of a response in this activity was 383 characters long and the minimum was 5, which indicates that some people do provide single word responses (which are still analyzable), but that this is the exception, not the norm. The shortest responses were provided in the crash test dummy question with a mean response length of 51 characters with a standard error of 3 characters. Typical responses were just one sentence long with two sentence responses being uncommon (there were 3). The shortest response was three words long. The coin and cylinder activity had a mean response length of 68 characters and the rest of the activities had mean response lengths that fell between 80 and 115 characters. In general, this paints a clear picture in which students provide responses that are generally long enough to provide appropriate physical details, but are not excessively verbose.

IV. RESULTS

The results of our analysis are summarized in Table II. Each row contains information relevant to the analyzed question response set from a particular lesson activity. The second column indicates the number of groups identified in the qualitative analysis of the response set. The third column indicates the number of responses in each response set. The fourth and fifth columns indicate the level of agreement between the human rater and the computer model when self-validation was performed, that is when the entire response set was coded using a model that was trained on the entire data set. The sixth and seventh columns indicate the level of agreement that was obtained when the response sets where randomly divided in half, two models were trained on the two half sets of data and each half of the data set was coded using the model trained on the other half.

As an additional check on our method of identifying groups an external rater with a background in physics and education was asked to code a random subset of the data chosen from four randomly selected activities: ball and track, car and coffee cup, hammer and feather, and ice skater and bowling ball. This amounted to coding a 12% sample of the data. The interrater agreement on these activities were 97%, 86%, 94%, and 98%, respectively. The lower agreement on the car and coffee cup activity is almost certainly due to the larger number of groups that emerged, which is indicative of less conceptual coherence within the data set. Responses in this set were more conceptually scattered, which makes analysis harder, for both human and computer. This issue is a key difficulty and is discussed in greater detail in subsection F, below. Still the agreement is reasonably good across all four activities, which suggests that while the groupings are not unique, they are distinctly recognizable.

In Table II we see five activities for which the automated assessment protocol was more successful and four for which it was less successful. The former are ball and track, crash test dummy, coin and cylinder, hammer and feather, and ice skater and bowling ball activities. All of which had self-validation match rates at or above 70% and cross-validation match rates at or above 60%. Four of them had cross-validation match rates at or above 70%. This corresponds to κ statistics that would be considered moderate to substantial by the scheme in Table I. The remaining activities had cross-validation match rates at or below 50%, and κ statistics that would be considered only fair by Landis and Koch rubric. Looking at the details of how the computer model grouped the responses to some of the questions in Table II provides insight into how best to write activities that might better yield response sets that lend themselves to this type of analysis. In the following sections we will look at these details for the five most successful cases and discuss the elements which likely led to lower quality results in the other four.

We should note that the agreement rate via self-validation is a good predictor of agreement rate obtained via cross validation. This is intuitive, but also useful to observe, because performing the self-validation without a cross validation may be a faster, easier way to assess the utility of a particular computer model in the developmental

TABLE II. Self-validation and cross-validation success rates for the 9 questions analyzed.

Question	Groups	Responses	Self-validated matches	κ	Cross-validated matches	κ
Ball and track	6	161	116 (72%)	0.6235	112 (70%)	0.5791
Car and coffee cup	14	154	75 (49%)	0.4206	55 (36%)	0.3614
Crash test dummy	6	161	136 (84%)	0.6535	135 (84%)	0.6743
Coin and cylinder	9	150	109 (73%)	0.6714	92 (61%)	0.5328
Beaker and coin	9	142	76 (54%)	0.4654	54 (38%)	0.3024
Hammer and feather	5	158	140 (89%)	0.7947	134 (85%)	0.7128
Train crash	13	105	52 (49%)	0.3888	38 (36%)	0.2282
Ice skater	8	110	77 (70%)	0.6298	77 (70%)	0.5653
Live and dead ball	9	89	49 (55%)	0.4396	44 (49%)	0.3321

stages of a system. Clearly, only cross validation with data that was not used in generating the model can provide certainty about the model's continued ability to classify new data, and there is no guarantee that future data will be well described by previously generated models. However, in this work we are interested in how the approach succeeds and fails, so in the succeeding sections we analyze the self-validation results because they provide a clearer lens on what is working without masking what is not.

The metrics for understanding the efficacy of the method category-by-category are precision and recall [52]. Precision and recall are based on the binary judgement of "relevant" and "irrelevant" and maximizing the return of relevant items and minimizing the return of irrelevant items [52]. In the context of our research the computer model would have coded a relevant response if it coded it as the human rater did, and it would have coded an irrelevant response if it coded it in another manner. Precision then characterizes the extent to which the computer model correctly rejects the irrelevant responses for each of the possible codes. Recall characterizes the extent to which the computer model correctly classifies all the possible relevant responses. We can think of precision as rejecting false positives and recall as obtaining true positives. It is clear that both are important for the success of the system since for any given code, a false positive corresponds to a missed true positive for another code.

A. Ball and track

In the question analyzed students were asked how the speed of a ball moving on a flat, low-friction track would change if it were measured at the end of a track that was twice as long. Fifty-eight of the 161 student responses indicated that the ball would move more slowly if the track were extended. A small, but interestingly large minority of 18 indicated that the ball's speed would increase. Fifteen responses were numeric in nature and 7 refused to choose between slowing down and maintaining the same velocity. Table III summarizes the number of responses identified by human coding as belonging to each group, the number matched that way by the computer, the total number the computer coded as belonging to each group, and the number mismatched. Naturally, the number matched plus the number mismatched must add to the number coded.

The most important thing to recognize is that it is clear in this question that students can only respond in a few reasonable ways. While the question may seem simplistic, the students' answers reveals something about how they are likely thinking about the physical system. This activity, which parallels Galileo's inertia experiments is, in part, designed to encourage students to consider the effect of friction, whether it is small, and how the relevant object would behave if it were zero.

This question and activity illustrates what we believe to be the key to writing questions and activities amenable to this type of automated analysis: questions should have a limited and small number of reasonable responses, and those responses should connect as directly as possible to distinct ways of thinking about the physical system. In this case we can see that a significant fraction of the students are saying that the ball will slow down. We can quite reasonably assume that they think this because of friction. This is not wrong, and it could be the starting point for a discussion about friction, and how it can mask Newton's first law, if we can successfully identify the students who provide that response. In this case, we can do that with better than 80% accuracy. We note that a small, but non-negligible fraction of the students indicated that the ball would speed up. There are multiple hypotheses about why they think this will happen, but ultimately it is necessary to ask a student why such a response was provided. This highlights an interesting facet of consideration for determining the minimum size of a response set for performing this type of analysis. We believe that it is not always more desirable to have more data to more accurately classify the dominant groups, but instead to have more examples of less frequent, but interesting responses to improve the matching rates for those types of responses.

B. Crash test dummy

In this question students were asked to explain why an unrestrained dummy, which is in motion during a car crash, ultimately stops. The best grouping scheme obtained is shown in Table IV. The majority of responses, 104, were not really an explanation, but simply a physical description

TABLE III. Group-by-group matching for the ball and track activity.

Group	Number	Matched	Recall	Coded	Mismatched	Precision
Slower	58	48	83%	68	20	71%
The same	43	39	91%	44	5	88%
Slower; not much	20	11	55%	16	5	69%
Greater	18	8	44%	14	6	57%
Numeric answer	15	9	60%	17	8	53%
Same or slower	7	1	14%	2	1	50%
Total	161	116	⋯	161	45	⋯

TABLE IV. Group-by-group matching for the crash test dummy activity.

Group	Number	Matched	Recall	Coded	Mismatched	Precision
Phys. Desc.	104	102	98%	111	9	92%
Force on dummy	34	31	91%	45	14	69%
Other force ideas	16	3	19%	6	3	50%
Newton's 3rd law	4	0	0%	2	2	0%
Momentum	2	0	0%	0	0	···
Acceleration	1	0	0%	1	1	0%
Total	161	136	···	161	25	···

of what had occurred in the video clip. Thirty-four responses focused on the idea that the dummy must feel a force to stop. Sixteen responses focused on force in different ways. Four responses cited Newton's 3rd law as the reason the dummy stops moving. Two explained the video clip in terms of the dummy's momentum being removed. One response cast the happenings purely in terms of acceleration. Table IV summarizes the group-by-group details for this activity in the same manner as Table III.

A small, but significant number of students indicated that the externally applied force is responsible for the change in the dummy's motion. The response set is dominated by answers that simply describe what is going on physically, that is, they do not provide any real explanation. While this may seem uninteresting at first, we observe that there is sufficient similarity across these responses that the computer correctly identified 98% of them with only 8% contamination from other responses. Thus, if a student answered the question by just describing what happened without explanation, the system could identify that, and provide feedback that focused on asking for a more detailed, explanatory answer. If the student correctly identified an external force as the source of the change in motion, the computer could identify that with high success as well. The more concerning problem is that a relatively high number of incorrect responses were inappropriately grouped into the force on dummy group. We must note, however, the small number of responses for training the model, which is likely responsible for this issue. It would be instructive to repeat the analysis after students had been given the feedback and asked to reanswer. A logical hypothesis is that the second set of answers would heavily populate the other groups, but it may be that additional groups emerge from that data set.

C. Coin stuck in a graduated cylinder

In this question students were asked to provide a method of obtaining a coin lightly stuck in the bottom of a graduated cylinder (so that they could not reach in and get it) and explain their method in the context of Newton's first law. The results from the best observed grouping scheme are summarized in Table V. Thirty-one responses suggested getting the coin simply by inverting the cylinder. Another 30 suggested hitting the bottom, 26 suggested hitting it on the table, 24 suggested that force was required. Ten suggested using gravity. Another 10 suggested multiple methods in a single response. Ten responses contained miscellaneous other ideas that were hard to classify with anything else expressed. Six responses focused on using the coin's inertia as a means of obtaining the coin. Three simply indicated shaking the cylinder with no physical explanation of why that would work. Table V summarizes the group-by-group details for this activity.

In the previous question we saw a good matching predominantly resulting from two dominant groups within the response set. The first thing we note in this example is that reasonably good matching is obtained despite a larger number of smaller groups. The first five groups are all matched reasonably well despite there being no one or two

TABLE V. Group-by-group matching for the coin stuck in a graduated cylinder activity.

Group	Number	Matched	Recall	Coded	Mismatched	Precision
Invert	31	31	100%	43	12	72%
Hit the bottom	30	21	70%	31	10	68%
Hit on table	26	18	69%	23	5	78%
Apply force	24	22	92%	29	7	76%
Use gravity	10	10	100%	10	0	100%
Mult. methods	10	1	10%	5	4	20%
Other ideas	10	3	30%	6	3	50%
Inertia ideas	6	2	33%	2	0	100%
Shake	3	1	33%	1	0	100%
Total	150	109	···	150	41	···

dominant responses. In this case, still, the smallest groups are not well matched but that will always be the case, and this again illustrates that the principle motivation for collecting more data is to improve the matching rates for these smaller groups.

This example also serves to illustrate how the clustering of responses is not unique, but the approach is useful despite this lack of uniqueness, and perhaps in some sense because of it. In this question students were asked to propose and explain a method for obtaining a coin. While grouping responses based on the proposed methods provides a distinct framework for classifying the responses that also works reasonably well for training computer models, another perspective for understanding the responses exists. Students frequently viewed the activity from the perspective of "a force must be applied to the coin" or "once the coin is in motion it will stay in motion." Because the activity focuses on Newton's first law, the latter perspective is more productive in explaining how to obtain the coin. We could group the responses based on this framework instead, but the variety of approaches that are consistent with either one of these perspectives makes it more difficult to train a computer model. Selecting the framework more suitable to automated analysis, however, does not preclude us, as intelligent actors, from realizing that other frameworks exist and deciding whether or not they warrant investigation.

D. Hammer and feather dropped on the moon

In this activity students were asked to predict (with justification), observe, and explain the behavior of a hammer and feather dropped on the moon. Watching the video reveals that, in the absence of air resistance, the hammer and feather fall at the same rate. The question we analyze asked students to predict the motion with explanation. In this question 98 responses indicated that the motion for the hammer and feather would be the same. Twenty-seven responses indicated that the hammer and feather would float when released. Twenty-two indicated that the hammer would fall faster. Five responses were really physical descriptions that did not make a clear prediction and six responses contained miscellaneous other ideas that were hard to group with anything else. Table VI summarizes the group-by-group details for this activity.

We can see that the majority of the students correctly predicted the behavior of the two objects and for the three major groupings, the computer model correctly matched the human rater more than 85% of the time. The contamination with other responses in those three groups was small, consistent with the overall high match rate of 89%. Interestingly, nearly 15% of the responses are consistent with failing to understand that there is gravity on the moon. While this finding is hardly earth shattering for the physics education community, it is important to know from an instructor's perspective, and illustrates how an online learning environment, such as ours, coupled with fast, automated feedback can be useful, in the context of a MOOC or a traditional class. While only 15% of students thought this, that is still a noteworthy fraction. In a lecture-class scenario that would be sufficient to warrant explicit mentioning in class. Many experienced teachers have anecdotal stories of being surprised by something their students did not know. This type of implementation could allow us to observe, quantify, and address these types of issues.

This question illustrates a potential benefit of fast feedback, as well. If the computer learning environment can correctly identify the students who hold a specific idea, or are using a specific approach to a problem, it can provide feedback to the student while they are still working on the problem, instead of later, once the instructor has had time to review the responses.

E. Ice skater throws a bowling ball

In this question students were asked which object, an ice skater or a bowling ball, experienced greater acceleration when the ice skater throws the ball, and moves backwards as a result. In this question 32 of the responses indicated that the bowling ball felt more acceleration because it had less mass. Twenty-four responses indicated that the bowling ball experienced greater acceleration because it had greater speed. Eighteen responses indicated the bowling ball had greater acceleration and assigned force as the reason. Ten responses indicated that the skater accelerated faster. Nine responses indicated that the bowling ball accelerated faster, and explained in terms of Newton's laws explicitly. Eight responses indicated that the skater and bowling ball experience the same acceleration, 4 responses suggested that the bowling ball accelerated

TABLE VI. Group-by-group matching for the hammer and feather on the moon activity.

Group	Number	Matched	Recall	Coded	Mismatched	Precision
Same	98	93	95%	101	8	92%
Float	27	23	85%	25	2	92%
Hammer is faster	22	20	91%	22	2	91%
Phys. Desc.	5	3	60%	6	3	50%
Misc. Ideas	6	1	17%	4	3	25%
Total	158	140	⋯	158	18	⋯

TABLE VII. Group-by-group matching for the ice skater and bowling ball (BB) activity.

Group	Number	Matched	Recall	Coded	Mismatched	Precision
BB/less mass	32	29	91%	35	6	83%
BB/greater speed	24	18	75%	23	5	72%
BB/feels force	18	11	61%	22	11	50%
Skater	10	4	40%	7	3	57%
BB/Newton's laws	9	5	56%	7	2	71%
Same	8	6	75%	6	0	100%
BB/other ideas	5	2	40%	5	3	40%
BB/moves further	4	2	50%	5	3	40%
Total	110	77	...	110	33	...

more rapidly and explained that in terms of the distance the ball moved. Five responses indicated that the bowling ball accelerated faster and justified that in terms of other miscellaneous ideas that did not fit with other ideas expressed. Table VII summarizes the group-by-group details for this activity.

This example again illustrates how the design of the activity may result in response sets that are more appropriate for this type of analysis. Because this video features only an ice skater and a bowling ball, only one of those can be chosen as having a greater acceleration. Nothing else makes sense. Furthermore, only a few physical explanations can be used to justify either choice. One can focus on the mass difference between the skater and bowling ball, one can focus on the apparent difference in speeds between the two objects as they begin their respective motions, etc.

F. Other activities

Though the match or mismatch picture yields enlightening results for some but not all of the activities, we can learn some things from some of the other activities as well. An activity that illustrates what can happen when there are too many ideas for students to focus on, or if too many groupings emerge is the car and coffee cup activity. This activity centers on a video in which an actor leaves his coffee cup on his car and then drives off. The coffee cup predictably falls straight down and students are asked to explain the motion in terms of Newton's first law. Students focused on a wide range of ideas: gravity, horizontal motion, vertical motion, whether the car and cup were attached, whether the cup felt a force, and others. This is evident from the 14 different groups that emerged from that response set. Another natural result of more groupings is that each group contains fewer responses and the computer has less data to train on. It would be interesting to see whether the computer could more effectively classify the responses with a significantly larger data set.

The activity that centers on the motion of the coin on the beaker parallels the coffee cup activity discussed above, and we observe the same problem: that students focused on a variety of ideas, many different groups emerged, which makes it difficult to achieve success with the automated assessment procedure. If a question results in a response set with a relatively small number of conceptually distinct responses then this analysis approach is more effective. If a question results in a response set that contains many, hard to distinguish ideas the analysis approach will likely not be effective.

In the train crash activity again we see a larger number of groupings emerge, and generally poorer matching between human and computer. That activity also suffers because the video contains two identical trains, and requires students to discuss the concept of force in that context. The logical result is that most responses focus on trains and force, but may convey very different ideas. This makes it difficult for the computer to resolve, via the presence or absence of other features, the subtle differences in the responses that convey to the human rater the differences in meaning that lead to different classification.

In the live and dead ball activity we also observed an example of a single idea being used in multiple ways. Virtually every response to this question contains the word force and constructs an explanation of what is going on using the concept of force. This makes for a response set that is very confusing for automated assessment. A relatively large number of groupings emerged for this question. However, in this case, the computer's inability to classify conceptually different responses is more likely due to students' use of the same words to convey different ideas. Potentially, three approaches could remedy this issue. The first, and in some sense easiest, is to take more data hoping that a better-trained model will better tag responses. The second is to require longer, more detailed responses which will allow the computer to have a larger body of text upon which to train. The third approach is to attempt to rewrite the activity to try to elicit students' ideas in ways that are less textually similar. At this point there is no way to know *a priori* whether any of these three methods will yield improvement; they must be tried to empirically judge.

V. DISCUSSION

In this work we present results from applying automated analysis techniques to student responses to short-answer questions. Further research in this direction is required

before it can be stated whether these results constitute what can typically be attained for short-answer questions when 100 or 200 responses are collected. As we will discuss in more detail below, we have seen examples of success that clearly warrant further investigation.

The first, most important metric for judging the success of this type of analysis is the rate of agreement between the trained computer model and the human rater. We care both about the κ statistic and the percent agreement because the κ statistic allows us to assess the effectiveness of the automated assessment while controlling for random agreement, which is a viable concern, but ultimately for implementation a high percent agreement is the better figure of merit. Exactly how high will be discussed in some detail at the end of this section. From Table II we can see that the best results exceed 70%, but the results are far from uniform.

Comparing the fourth and sixth columns (as well as the fifth and seventh) of Table II shows relatively good agreement between the performance of a self-validation on an entire response set and the performance of cross validation on each half of the data set. This may be of interest to researchers, as an indicator that performing the self-validation on a model is typically a good predictor of how well the cross validation results will be.

Generally, our analysis has suggested to us that within a given response set larger groups are more likely to be correctly matched by the computer models. This is quite reasonable, and generally expected. At the same time, it is useful to explicitly see that expectation borne out. A rough criterion of twenty responses within a group appears to be a good guideline, though this is currently just an estimate. It is likely that this is a desirable condition but insufficient to guarantee good classification. Again, our current goals are to explore the potential utility of this approach and identify strategies that maximize the likelihood of success in applying the approach; we are not yet in a position to outline concrete rules for successful application. With the rough guideline of about twenty responses per group in mind, and the observation of five to ten groups emerging in questions that do work well, we can also identify a good general guideline for the minimum size of a data set for which we might reasonably expect this approach to work. That guideline is between 100 and 200 responses, depending on the number of groups that are identified in the response set. From this estimate we can see that our data sets are very much at the lower limit of what is feasible. It is nonetheless interesting to observe some success with response sets of this size, and our results suggest that there is no reason to *a priori* reject this approach for data sets of this size. At the same time, larger data sets being more desirable suggests that collaborations that could combine smaller response sets may present a significant advantage in further exploring machine learning in physics education.

In order to interpret our results in a more meaningful way it is clear that the standard for comparison emerges from the question: How often can our tutoring system misconstrue what a student says and still provide the student with functional support? In that regard human tutors clearly set the standard. Our goal should be to create a system that approaches the same accuracy that a human tutor would in assessing students' work. In some sense this is an ambitious goal, since the human tutor has judgment and intelligence, but it is intuitive that even human tutors sometimes make mistakes in assessing students' understanding. An electronic system that approaches that error rate would have potential as an online instructional tool that could provide more interactive feedback to students than online learning systems currently provide while failing only as much as a human tutor would anyway. It is tempting to think of human tutors as nearly flawless, but research suggests otherwise. Research suggests that tutors with good content knowledge and limited tutoring experience (as is typical of real tutors at the high school and college levels) may correctly assess student understanding as little as 70% of the time [21]. With a 30% failure rate, several of the questions we have analyzed are already getting close to what a human tutor can do.

Moreover, our results point us in a clear direction towards building activities that promote a more interactive system. Just as our system currently provides video responses to students' typed questions, it is possible with continued development effort to provide video feedback based on student responses to appropriate lesson questions.

While the results suggest some success with data taken so far, an important question for ultimate implementation is one of reproducibility. We must keep in mind the possibility that if we presented our activities to another group of students we would obtain different data. If the data are slightly different, then the groups that emerged from these data and models trained on them would likely be very useful in analyzing that data. If the data are very different then this would not be the case. In the context of this research the idea of saturation is important, and it is not easy to say *a priori* when saturation is likely to occur [53]. It also cannot safely be assumed that student populations will not change with time, resulting in differences in data sets. These possibilities suggest possible impediments to further development and implementation of automated assessment of short-answer responses, but they are impediments that can only be understood through further research.

In considering the broader implications of this research it is useful to note that automated assessment, such as we have done, does not have to be done in the context of online tutoring systems. This type of scheme could be used in any application where moderately large quantities of short text responses are collected. Clearly the easiest method of collecting these responses would be through online means, but it is not the only one. One of the implications of this research is that an instructor who can collect a hundred or more text responses to a question can, in principle, perform this kind of analysis using the software, which is freely

available for download. This of course requires a commitment to learning how to best apply the software to these types of data sets. While this alone does not promise that the methods will work for all data sets, wider attempts to apply these types of approaches should help the community understand to what extent they do work, and when they can be productively applied in instruction and assessment.

Another consideration of importance in this research is sample diversity, and, in particular, the diverse linguistic expression present in the United States, and, indeed, the world. Our sample was drawn from students in North Central Kansas, and the sample likely does not completely represent the national population on multiple metrics. As noted previously, English language learners are likely not represented in the sample. Regional differences in patterns of linguistic expression may also manifest in written communication, and our study cannot resolve such differences or provide information about their potential impact on applying the natural language analysis. Additional research is needed to investigate these topics. A logical extension of this recognition is that the approach could be applied beyond English, and understanding that potential is an important and interesting direction of study.

VI. CONCLUSIONS

In this paper we have presented evidence for success in the automated analysis of short text responses to conceptual questions in the context of an online tutoring system. We have demonstrated that this line of inquiry is interesting because it has the potential to allow our synthetic tutor to respond to students based on their answers to questions, which in turn allows a greater level of interactivity and could allow our system to better fulfill the roles played by real tutors. Beyond our specific project, this analysis approach has important implications for online homework systems, and other forms of online instruction, such as MOOCs, that could benefit from automatic assessment of students' typed text.

These results represent our initial attempts at applying this analysis scheme to these type of data. The agreement between human and computer is far better than random chance, but worse than human-human agreement, which is not surprising. There is evidence that our best results are beginning to show human-computer agreement that may be good enough to rival real human tutor's abilities to accurately assess student understanding. However, the worst results are far from this good. Using larger data sets will almost certainly improve the agreement rates, but at the cost of additional analysis time, for an approach that is already time consuming. Additional research is clearly needed to assess the benefits and costs associated with data sets of different sizes.

The fact that not all questions seem to readily generate response sets that are readily clustered, even by humans, is problematic. However, we have discussed methods of writing questions and activities that are likely to produce response sets that are more appropriate for this type of analysis, which focus on providing students with small discrete numbers of physical objects and physics concepts that can be put together to construct explanations of physical phenomena. While one could adopt a view that this is nothing more than a path towards multiple choice, we argue that this view overlooks the importance of giving students the opportunity to express their ideas about the behavior of physical systems in their own words. Results like those we present here may ultimately allow us to assess students'responses to conceptual questions in a manner that is not unlike how online homework systems have allowed us to efficiently assess students'answers to mathematical problems. We believe that this type of approach has tremendous promise as a similar tool in online assessment.

At the same time, there is much research work that still must be done in this area. Further investigations extending our work across more physics content would be beneficial from a pedagogical perspective, but would also serve as a test to determine how effective our strategies for developing lesson materials actually are. Using these activities with more groups of students will allow us to have greater confidence that the groupings we have established with this data are useful for analyzing future data. One of the most time-consuming facets of this research is the manual analysis of data that is used for training the computer models. Fully automated data-mining techniques for text analysis have been developed [3,54]. Prior to performing this analysis, preliminary tests with fully automatic assessment were performed. Our initial tests did not produce conceptually meaningful clusters, but further, systematic, investigation in this direction is necessary. If successful, this kind of approach may relieve some of this burden and allow for faster training of computer models. That in turn could make the idea of automated analysis more practical for a broader range of instructors and researchers. Continuing this line of investigation will provide the best chance of exploiting the tools of text analysis and data mining to provide more interactive physics learning environments that are suitable for today's educational requirements, and modern, web-savvy students.

ACKNOWLEDGMENTS

We acknowledge the support of the United States National Science Foundation under Grants No. REC-0632587 and No. REC-0921628. We also acknowledge Bryan Maher and Josh Gross for their work on the Pathway Active Learning Environment. We thank N. Sanjay Rebello, Kimberly Staples, Penny Blue, and Sarah Nuss-Warren for useful discussions about the Pathway Active Learning Environment's functionality. This publication was financially supported in part by the Kansas State University Open Access Publishing Fund.

[1] I. E. Allen and J. Seaman, *Going the Distance: Online Education in the United States, 2011* (Babson Survey Research Group, Babson Park, MA, 2011).

[2] R. S. Baker and Y. Kalina, The state of educational data mining in 2009: A review and future visions, *JEDM-J. Edu. Data Mining* **1**, 3 (2009).

[3] C. Romero and S. Ventura, Educational data mining: A review of the state of the art, IEEE Trans. Man Cybern. C **40**, 601 (2010).

[4] R. Carlson, K. Genin, M. Rau, and R. Scheines, Student profiling from tutoring system log data: When do multiple graphical representations matter? in *Proceedings of the 6th International Conference on Educational Data Mining*, edited by S. K. D'Mello, R. A. Calvo, and A. Olney (International Educational Data Mining Society, Worcester, MA, 2013), p. 12.

[5] K. E. Boyer and A. Ezen-Can, Unsupervised classification of student dialogue acts query-likelihood? in *Proceedings of the 6th International Conference on Educational Data Mining*, edited by S. K. D'Mello, R. A. Calvo, and A. Olney (International Educational Data Mining Society, Worcester, MA, 2013), p. 20.

[6] E. Snow, G. T. Jackson, L. Varner, and D. S. McNamara, Investigating the effects of off-task personalization on system performance and attitudes within a game-based environment, in *Proceedings of the 6th International Conference on Educational Data Mining*, edited by S. K. D'Mello, R. A. Calvo, and A. Olney (International Educational Data Mining Society, Worcester, MA, 2013), p. 272.

[7] A. T. Corbett, K. R. Koedinger, and J. R. Anderson, Intelligent tutoring systems, *Handbook of Human Computer Interaction* (Elsevier, Amsterdam, 1997), p. 849.

[8] F. Reif and L. A. Scott, Teaching scientific thinking skills: Students and computers coaching each other, Am. J. Phys. **67**, 819 (1999).

[9] K. Van Lehn, C. Lynch, K. Schulze, J. A. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill, The andes physics tutoring system: Lessons learned, *Int. J. Artif. Intell. Educ.* **15**, 147 (2005).

[10] A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney, Autotutor: An intelligent tutoring system with mixed-initiative dialogue, IEEE Trans. Ed. **48**, 612 (2005).

[11] D. T. Seaton, Y. Bergner, I. Chuang, P. Mitros, and D. E. Pritchard, Who does what in a massive open online course? Int. J. Hum-Comput. St. **68**, 223 (2010).

[12] S. Cooper and M. Sahami, Reflections on stanford's moocs, Commun. ACM **56**, 28 (2013).

[13] S. W. Bonham, R. Beichner, and D. Deardorff, Online homework: Does it make a difference? Phys. Teach. **39**, 293 (2001).

[14] K. K. Cheng, B. A. Thacker, R. L. Cardenas, and C. Crouch, Using an online homework system enhances students' learning of physics concepts in an introductory physics course, Am. J. Phys. **72**, 1447 (2004).

[15] B. S. Bloom, The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring, Educ. Res. **13**, 4 (1984).

[16] P. A. Cohen, J. A. Kulik, and C. C. Kulik, Educational outcomes of tutoring: A meta-analysis of findings, Am. Educ. Res. J. **19**, 237 (1982).

[17] D. Marinelli and S. Stevens, Synthetic interviews: The art of creating a 'dyad' between humans and machine-based characters, in *Interactive Voice Technology for Telecommunications Applications* (IEEE, Piscataway, NJ, 1998), pp. 43–48.

[18] D. Wood, J. S. Bruner, and G. Ross, The role of tutoring in problem solving*, J. Child Psychol. Psychiatry **17**, 89 (1976).

[19] A. C. Graesser and N. K. Person, Question asking during tutoring, Am. Educ. Res. J. **31**, 104 (1994).

[20] C. Juel, What makes literacy tutoring effective? Read. Res. Q. **31**, 268 (1996).

[21] M. T. H. Chi, S. A. Siller, and H. Jeong, Cognit. Instr. **22**, 363 (2004).

[22] E. B. Page, The imminence of grading essays by computer, *Phi Delta Kappan* **47**, 238 (1966).

[23] M. A. Hearst, The debate on automated essay grading, IEEE Intell. Syst. Appl. **15**, 22 (2000).

[24] S. Dikli, An overview of automated scoring of essays, *J. Tech. Learn. Assess.* **5**, 1 (2006).

[25] R. H. Nehm and M. Ha, Item feature effects in evolution assessment, J. Res. Sci. Teach. **48**, 237 (2011).

[26] C. Leacock and M. Chodorow, C-rater: Scoring of short-answer questions, Comput. Hum. **37**, 389 (2003).

[27] S. G. Pulman and J. Z. Sukkarieh, Automatic short answer marking, in association for computational linguistics, in *Proceedings of Second Workshop on Building Educational Applications Using NLP, Ann Arbor, Michigan* (Association for Computational Linguistics, New Brunswick, NJ, 2005), p. 9.

[28] S. Jordan, Student engagement with assessment and feedback: some lessons from short-answer free-text e-assessment questions, Comput. Educ. **58**, 818 (2012).

[29] M. Ha, R. H. Nehm, M. Urban-Lurain, and J. E. Merrill, Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations, CBE Life Sci. Educ. **10**, 379 (2011).

[30] E. P. Beggrow, M. Ha, R. H. Nehm, D. Pearl, and W. J. Boone, Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance? J. Sci. Educ. Technol. **23**, 160 (2014).

[31] R. M. Dawes, D. Faust, and P. E. Meehl, Clinical versus actuarial judgment, Science **243**, 1668 (1989).

[32] J. B. Brungardt and D. Zollman, Influence of interactive videodisc instruction using simultaneous-time analysis on kinematics graphing skills of high school physics students, J. Res. Sci. Teach. **32**, 855 (1995).

[33] L. T. Escalada and D. A. Zollman, An investigation on effects of using interactive digital video in a physics classroom on student learning and attitudes, J. Res. Sci. Teach. **34**, 467 (1997).

[34] R. J. Beichner, The impact of video motion analysis on kinematics graph interpretation skills, Am. J. Phys. **64**, 1272 (1996).

[35] P. Laws and H. Pfister, Using digital video analysis in introductory mechanics projects, Phys. Teach. **36**, 282 (1998).

[36] D. Brown and A. J. Cox, Innovative uses of video analysis, Phys. Teach. **47**, 145 (2009).

[37] R. Karplus and D. P. Butts, Science teaching and the development of reasoning, J. Res. Sci. Teach. **14**, 169 (1977).

[38] A. Champagne, L. Klopfer, and J. Anderson, Factors influencing the learning of classical mechanics, Am. J. Phys. **48**, 1074 (1980).

[39] R. T. White and R. F. Gunstone, *Probing Understanding* (Falmer Press, Great Britain, 1992).

[40] R. E. Scherr, P. S. Shaffer, and S. Vokos, The challenge of changing deeply held student beliefs about the relativity of simultaneity, Am. J. Phys. **70**, 1238 (2002).

[41] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques* (Elsevier, New York, 2011).

[42] C. Cortes and V. Vapnik, Support-vector networks, Mach. Learn. **20**, 273 (1995).

[43] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods (Cambridge University Press, Cambridge, England, 2000).

[44] P. Domingos and M. Pazzani, On the optimality of the simple bayesian classifier under zero-one loss, Mach. Learn. **29**, 103 (1997).

[45] S. Kim, K. Han, H. Rim, and S. H. Myaeng, Some effective techniques for naive Bayes text classification, IEEE transactions on knowledge and data engineering **18**, 1457 (2006).

[46] C. Rosé, Y. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger, and F. Fischer, Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning, Int. J. Comput. Supp. Collab. Learn. **3**, 237 (2008).

[47] Lightside researcher's workbench, http://ankara.lti.cs.cmu.edu/side/download.html.

[48] K. C. Haudek, J. J. Kaplan, J. Knight, T. Long, J. Merrill, A. Munn, R. Nehm, M. Smith, and M. Urban-Lurain, Harnessing technology to improve formative assessment of student conceptions in stem: Forging a national network, CBE Life Sci. Educ. **10**, 149 (2011).

[49] R. H. Nehm, M. Ha, and E. Mayfield, Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations, J. Sci. Educ. Technol. **21**, 183 (2012).

[50] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. **20**, 37 (1960).

[51] J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data, Biometrics **33**, 159 (1977).

[52] M. K. Buckland and F. C. Gey, The relationship between recall and precision, J. Am. Soc. Inf. Sci. **45**, 12 (1994).

[53] G. Guest, A. Bunce, and L. Johnson, How many interviews are enough? An experiment with data saturation and variability, Field Method. **18**, 59 (2006).

[54] P. Berkhin, *A survey of clustering data mining techniques, in Grouping Multidimensional Data*, edited by J. Kogan, C. Nicholas, and M. Teboulle (Springer, Berlin, Heidelberg, 2006), pp. 25–71.