Article

Educational and Psychological Measurement
2014, Vol. 74(6) 950–974

DOI: 10.1177/0013164414521634
epm.sagepub.com
SAGE

Automated Scoring of Teachers' Open-Ended Responses to Video Prompts: Bringing the Classroom-Video-Analysis Assessment to Scale

Nicole B. Kersting[1], Bruce L. Sherin[2], and James W. Stigler[3]

Abstract

In this study, we explored the potential for machine scoring of short written responses to the Classroom-Video-Analysis (CVA) assessment, which is designed to measure teachers' *usable* mathematics teaching knowledge. We created naïve Bayes classifiers for CVA scales assessing three different topic areas and compared computer-generated scores to those assigned by trained raters. Using cross-validation techniques, average correlations between rater- and computer-generated total scores exceeded .85 for each assessment, providing some evidence for convergent validity of machine scores. These correlations remained moderate to large when we controlled for length of response. Machine scores exhibited internal consistency, which we view as a measure of reliability. Finally, correlations between machine scores and another measure of teacher knowledge were close in size to those observed for human scores, providing further evidence for the validity of machine scores. Findings from this study suggest that machine learning techniques hold promise for automating scoring of the CVA.

[1]University of Arizona, Tucson, AZ, USA
[2]Northwestern University, Evanston, IL, USA
[3]University of California Los Angeles, CA, USA

Corresponding Author:
Nicole B. Kersting, Department of Teaching, Learning, and Sociocultural Studies, University of Arizona, Tucson, AZ 85721, USA.
Email: nickik@email.arizona.edu

Keywords

classroom-video-analysis assessment, automated scoring, naïve Bayes, teacher knowledge, short answer items, text classification

The classroom-video-analysis (CVA) instrument is an assessment designed to measure teachers' *usable* teaching knowledge, that is, the knowledge that mathematics teachers are able to access and use in a classroom situation. Initial empirical evidence for the topic of fractions suggests that CVA scores, which are based on teachers' written analyses of short video clips of classroom instruction, are reliable and valid, predicting instructional quality and student learning (Authors, 2008; Authors, 2010; Authors, 2012). This is encouraging news for a field in which measures that predict outcomes of interest have been hard to come by (Darling-Hammond & Baratz-Snowden, 2007; Hill & Ball, 2004; Hill, Schilling, & Ball, 2004). Despite the promising results, however, wider use of the CVA scales has been limited due to the costly and labor-intensive process required to score teachers' written responses (Hill, Ball, Sleep, & Lewis, 2007).

In the current study, we explored the feasibility of machine scoring as an alternative method of scoring teacher responses. Using human-scored responses from three different CVA assessments (on fractions, ratio and proportions, and on variables, expressions and equations), we created naïve Bayes classifiers. The behavior of the resulting classifiers was then examined in multiple ways. First, we compared the scores produced by these classifiers to those that were assigned by human raters. Second, we used cross-validation techniques (explained below) to explore the generalizability of the performance our classifiers. Third, we compared the size of correlations between human and computer scores to correlations between human scores and the length of responses. In this context, the correlation with length-of-response provided a lower bound performance criterion for our classifiers. Fourth, we computed the internal consistency of both human and machine scores. Finally, to further explore the validity of the computer-generated scores, we related them to scores from another measure of teacher knowledge, the Mathematics Knowledge for Teaching instrument. The strength of association was then compared to the strength of association for human-generated scores.

It is not our assumption that, to be useful, machine-generated scores must produce results that are always completely in alignment with human scores. Indeed, one of the most intriguing possibilities is that machine-generated scoring might produce results that complement those produced by human scorers (Attali, 2011, 2013). This requires, however, that we have a clear understanding of the behavior of our machine classifiers, and how that behavior is similar and different from that of human raters. The above analyses provide that assessment, and it is that examination that is the heart of the work reported herein.

The CVA Approach to Measuring Usable Teaching Knowledge

Teaching is a complex activity that requires teachers to interpret classroom events in real time in order to make instructional decisions that benefit student learning. This

requires not only that teachers have many different kinds of knowledge but also that they are able to access and use their knowledge in the classroom. The CVA approach is focused on measuring this *usable* knowledge because it is the knowledge most likely to directly affect instruction and student learning. To approximate as much as possible a real classroom situation, the approach uses authentic video clips of mathematics instruction, presented online, that teachers are asked to view and then analyze in writing.

Under the CVA approach teachers' short, written analyses of the observed teaching events are taken to reflect their usable teaching knowledge and scored according to four different, yet related rubrics that address key aspects of instruction (Cohen, Raudenbush, & Ball, 2003; Pauli & Reusser, 2011). Teachers who are able to produce more sophisticated analyses of the observed teaching episodes and who thus receive higher scores have greater usable teaching knowledge. CVA items are not designed to assess narrow and prespecified content knowledge, and there are no expected right or wrong answers in the traditional sense. This is in contrast to more typical short-answer items that focus on assessing particular content knowledge and are designed to elicit a fairly closed-ended correct answer (Brew & Leacock, 2013).

Recent Progress in Automated Text Scoring

Over the past decade considerable progress has been made in the development and application of automated text analysis techniques for scoring of written and spoken text, with much of the work being focused on essays (Shermis & Burstein, 2013; Shermis, Burstein, Higgins, & Zechner, 2010). Current systems can produce scores more quickly and reliably and at a lower cost than trained human raters (Topol, Olson, & Roeber, 2010). A growing number of studies have demonstrated close agreement between human- and machine-generated scores (Shermis & Burstein, 2013). However, some caution is necessary in interpreting these results; a simple stance, in which machine scorers are understood as merely replicating the work of human raters, is not justified (Attali, 2013).

Automated scoring is currently being employed in real-world niches that capitalize on its strengths while minimizing the impact of its weaknesses. Computer-generated scores are now routinely used to score essays in low-stakes testing situations or to provide instructional feedback (Shermis & Hammer, 2012). In contrast, high-stakes tests such as the SAT, GRE, or GMAT use computer-generated scores in conjunction with human–rater-assigned scores, either by combining both kinds of scores (essentially treating the computer algorithm as a second rater) or by using machine scores as a control for human scoring (Zhang, 2013).

The success of automated scoring has been found to depend on the nature of the material, the amount of text that can be analyzed to determine the score, and the features to be scored. In a large-scale study by Shermis and Hammer (2012), computer algorithms were found capable of producing essay scores that closely corresponded to rater-assigned scores. For source-based essays (i.e., essays based on specific

information made available at the time of writing) computer scoring algorithms outperformed human raters, though automated scoring was slightly less accurate for free-form essays. Similarly, human scores on rubrics that address overall style, quality, and correctness of an argument or character development showed lower agreement with machine scores, presumably because they rely more heavily on understanding the meaning of the written text, and less on rubrics that address structural features of the composition, grammatical correctness, or word choice (Attali, 2013; Shermis & Hammer, 2012).

Less is known about the feasibility of machine scoring for short answer items. Short answer items pose a different set of challenges than essays for the obvious reason that the amount of text that can be analyzed for scoring is small—usually not more than a few sentences. Furthermore, this challenge is likely to be heightened for certain types of short-answer items, particularly those that are more open-ended in nature. It is perhaps for this reason that much of the work on short answer items has concentrated on items that assess content knowledge (Brew & Leacock, 2013). In these cases, a particular type of analysis that makes use of Latent Semantic Analysis (LSA) has been the method of choice. First, the concepts and ideas that represent a correct response can be captured in terms of one or more ''model'' or ''best'' answers, usually specified a priori during the item writing phase (Brew & Leacock, 2013). Responses are then scored by essentially comparing them, using LSA, to the model answers. LSA is a vector space method that can be used to evaluate the similarity of passages of text. It is thus most useful when model answers exist to which a response can be compared (Berry, Dumais, & O'Brien, 1995; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Foltz, Gilliam, & Kendall, 2000; Foltz, Laham, & Landauer, 1999; Landauer, 2003).

High correlations between machine- and human-assigned scores using LSA (ranging from $r = .88$ to $r = .97$) have been reported for a variety of item formats that are more closed-ended in nature, including written and spoken summaries, research syntheses, and short answers, indicating greater agreement between computer and trained rater than between any pair of trained raters (Streeter, Bernstein, Foltz, & DeLand, 2011). Streeter et al. (2011) sum up this situation by observing that

> as with human scoring, the accuracy of automated scoring depends on several factors, including task clarity and well-designed training data. For automated scoring systems, the degree of constraint expected in the constructed responses is somewhat more important than it is with human ratings. (p. 3)

Taken together, these examples indicate clearly that good results can be obtained with the proper resources and expertise both in the instrument and the classifier development but that the results depend on the nature of the material to be scored and the type of rubric to be applied. Hence, a central question of this work is whether teachers' short, written analyses in CVA assessments are suitable for automated scoring, and whether the CVA rubrics are of the sort amenable to automation. Although we did not find any studies that specifically explored automated scoring of short

answers in response to video clips with an associated prompt, both the short answer item format and the source-like nature of the video clips appeared to hold promise for developing well-functioning scoring classifiers, whereas the relative open-endedness of the prompt might introduce challenges for automation, including the particular classification techniques used.

Overview of the Approach

To automate the scoring of open-ended text responses, we leveraged techniques from computational linguistics and machine learning. As suggested above, an approach that one might employ would be to craft model answers—both good and bad—and then to compare responses to those models. Given the open nature of the CVA tasks, we judged that this was not a profitable direction in which to proceed. Instead, we used an approach that can learn from human raters, while itself making few assumptions about what makes a good or bad response.

To do so, we conceptualized the scoring of responses as a problem of supervised text classification (Sebastiani, 2002). Supervised text classification algorithms are used, for example, to label web pages according to a set of thematic categories (e.g., as being about sports, politics, or health) or to label e-mail messages as ''spam'' or ''not spam.'' Our goal in this case is to classify (or ''label'') text samples into the scoring categories that are specified by our rubrics.

Supervised text classification starts with a set of reference texts that have previously been labeled. This reference set is used to train the automated classifier, which can then be employed to label previously unlabeled responses. In our case, where the labels are scores, we started with a set of teacher responses that have been scored manually by trained raters.

Given this specification of our task, there exists a wide range of techniques that could be employed. Although most of these approaches are based solely on analyses of word frequencies in the text to be classified, the way those frequencies are used can vary greatly. Typically, a choice of technique is made both heuristically (which techniques seem likely to work based on theoretical analysis as well as previous experience with similar problems) and empirically (which techniques give the best results in the current situation). To further narrow the space of possible choices, we decided to start as simply as possible, adding more complex methods only when they appeared to offer a clear advantage. This approach makes it possible for us to better understand why and how our classifiers give the results that they do. Starting simple also provides a baseline against which the performance of more complex methods could be evaluated.

Weighing these considerations led us to settle on the creation of *naïve Bayes* classifiers. A naïve Bayes approach (Lewis, 1998; Zhang, 2004), such as the one we employ here, is based on statistical associations between text classifications and word frequencies, which are collected from a human-scored reference set. Naïve Bayes has the benefit that, in comparison with discriminative approaches such as logistic

regression, it has been shown to perform well with relatively small amounts of data, as is the case in the present work (Ng & Jordan, 2002). Furthermore, some of our early efforts that made use of naïve Bayes showed promise. Finally, as we show later in this article, naïve Bayes has a conceptual simplicity that makes it possible understand, in some numerical detail, precisely how the automated classifiers arrive at labels for a specific case.

Like all these methods of automated classification, naïve Bayes cannot be used ''off-the-shelf'' without some exploration and adjustment. In addition, once an approach is selected, some tinkering is required to fine-tune the analysis. For example, it is not uncommon to find that removal of certain words prior to classification may improve scoring accuracy.

In this study, we used automated text analysis algorithms to address the following research questions:

1. *How similar are machine scores and human scores?* We used Cohen's quadratic weighted Kappa and measures of correlation to indicate agreement and correspondence between computer-produced and rater-assigned scores. Although it is important to investigate the similarity between human and machine scores, we do not interpret this similarity as implying that we must think of machine scores as *replicating* human scores.
2. *Do machine scores exhibit internal consistency?* Treating clips as items, we computed Cronbach's alpha as a measure of internal consistency for machine and human scores. We viewed this as a measure of one type of reliability.
3. *Is there evidence for criterion-related validity of the machine scores?* We examine if the relationship between computer-generated scores and an external criterion (another measure of teacher knowledge) is comparable to the relationship observed between that criterion and human scores.

Method

Instrument Description

We used data from three different CVA assessments. We analyzed responses from 238 mathematics teachers who had analyzed 13 video clips of the CVA fraction assessment, another 238 teachers who had completed the ratio and proportion (RP) assessment also consisting of 13 video clips, and 249 teachers who had responded to 14 video clips of the CVA assessment on variables, expressions, and equations (VEE). These data were collected as part of a larger instrument development study, and the samples might be considered relatively small for developing automated classifiers.

To elicit teachers' usable teaching knowledge, we showed them a series of video clips of actual classroom situations and then asked them to ''analyze how the teacher and the student(s) interacted around the mathematical content.'' Teachers typed their responses into a web form. Although the analysis prompt was originally not designed

with automated scoring in mind, we had aimed to focus teachers' responses on key aspects of the teaching and learning process. We intentionally did not try to constrain which particular aspects in the video clips teachers might address because we expected those to vary depending on teacher knowledge.

We rated teachers' responses to the video clips according to four different, yet related rubrics, each consisting of three categories. The rubrics indicated the degree to which each response: (a) analyzed the mathematical content (MC) and (b) student thinking (ST) depicted in the video clip, included (c) suggestions for improving (SI) the observed teaching episode, and (d) analyzed the observed teaching episode in depth (DI). A response was rated as falling in the lowest category (a score of 0) when it did not refer to the mathematical content or student thinking, or did not provide any suggestions for improvement. A score of zero on the depth of interpretation rubric (DI) was assigned when the response provided a descriptive account of the clip without interpretation, or if the response contained broad judgments that were not substantiated.

A response was rated as falling in the middle category (a score of 1) when the mathematical content was addressed in descriptive ways (e.g., by referring to the particular mathematical problem at hand), when student thinking was addressed either by providing assessments of student thinking that might be substantiated by observed student actions (e.g., I don't think the student understood because every time the teacher asked a question he did not answer until the teacher reworded the question in such way that the correct answer was obvious) or when general pedagogical suggestions for improving the teaching episode were provided (e.g., the teacher should have waited longer to give the student the opportunity to think about his answer). For the DI rubric, a score of 1 was assigned when the response contained some interpretation, for example, in the form of a substantiated judgment or included several unconnected interpretive points.

A response was rated as falling in the highest category (a score of 2), when the mathematical content depicted in the video clip was analyzed in depth by extending the mathematics beyond what was shown in the video clip (teachers did not receive any credit for incorrect math), when student thinking was analyzed in direct relation to the mathematics being worked on, and when suggestions for improvement were based on the specific mathematics of the teaching situation. Finally, the highest score on the DI rubric was assigned when teachers provided an in-depth analysis that connected different interpretative points to form a coherent argument.

As part of the original instrument development study, interrater reliability was evaluated before scoring began and at midpoint to control for rater drift. Initial interrater reliability estimates ranged from 79% to 91% direct agreement between each rater and a master, with midpoint estimates computed between pairs of raters being close. The values suggest that raters assigned scores to teacher responses with a reasonable degree of consistency. Having good interrater agreement is an important prerequisite to develop well-functioning automated classifiers because subsets of the manually scored responses serve as training texts in the classifier development.

In Table 1, we provide two authentic teacher responses to one video clip and explain their scoring based on the four different rubrics. Later, we use the first example response to illustrate how the automated scoring works in detail. In the video clip the teacher assists a student during an independent work phase in which students create several equivalent fractions for a given fraction. The teacher reminds the student of the example they had worked out together as a class, pointing out that they had used the ''Giant 1'' (i.e., a fractional representation of 1) to either multiply or divide. After creating a few equivalent fractions together, the teacher leaves the student to finish the work.

Data Description

Across all three topic areas and all four rubrics, scores of ''0'' and ''1'' were much more frequent (between 80% and 90% of all scores) than scores of ''2'' (between 7% and 12%, depending on topic and rubric). The score distributions reflect our rubric constructions, where scores of 2 intentionally identify teachers with greater expertise.

As shown in Table 2, for two of the rubrics—mathematical content and depth of interpretation—the distributions of scores were similar and fairly stable across the three topic areas: approximately 45% of responses were scored as ''0,'' 45% as ''1,'' and 10% as ''2.'' Score distributions of the suggestions for improvement rubric were also fairly stable across topic areas. About two thirds of teacher responses did not include any suggestions for improvement, roughly 20% included general pedagogical suggestions, and another 10% made very specific mathematically based suggestions.

For the student thinking rubric, the distribution of scores differed by topic area. For the Fraction assessment, the score distribution was similar to those observed for the other rubrics, but for the RP and VEE assessments there were almost twice as many 0s than there were 1s and only a small percentage of responses were scored as ''2.'' One possible explanation for these differences is that teachers were more familiar with student thinking and understanding around fractions and hence better able to analyze student thinking in the fraction video clips, whereas their knowledge of student thinking about RPs and VEE was less developed. Across all topics, few teachers analyzed student thinking within the specific context of the mathematics shown in the clip.

These patterns generally held when we examined score distributions by clip, but there was also some variation among clips. For each of the topic areas there were several clips that produced relatively more 0s and fewer 1s and 2s (e.g., Clip 3 for VEE, or Clip 3 for RP), possibly a sign that those clips were more difficult to analyze, or perhaps just less engaging to teachers. Other clips appeared easier to analyze, producing relatively more 1s and 2s (e.g., Clip 8 of RP).

Based on the way we defined our scoring rubrics, we had fewer manually scored reference texts for some rubrics and rubric categories than others, which might affect the performance of our classifiers.

Table 1. Two Example Responses and Their Scoring on the Four Rubrics.

Example 1	Rater-assigned scores
"The teacher is asking the student questions to narrow down his search to find equivalent fractions. I don't know if the boy understood the meaning of the giant 1. It was almost like the student was guessing, and judging how the teacher responded he would know if he was right or wrong."	*Mathematical Content*: 1 Because the response descriptively refers to the mathematical task (finding equivalent fractions using the giant 1) but not further analyzes the math beyond what was directly observable in the clip *Student Thinking*: 1 There is explicit concern for student thinking (I don't know if the boy understood) that is substantiated (guessing and using teacher response to gauge if right or wrong). *Suggestions for Improvement*: 0 No suggestion is included in the response. *Depth of Interpretation*: 1 Substantiated judgment.
Example 2	Rater-assigned scores
"I liked the idea of multiplying or dividing by a 'giant 1' when finding equivalent fractions. I do think the teacher could have given the student a little more 'think time' before repeating her questions—she was tapping her fingers and seemed distracted and like she was rushing the student a bit in his thinking."	*Mathematical Content*: 1 Because the response only refers to mathematical content shown in the video clip in descriptive ways, finding equivalent fractions by multiplying or dividing by a "giant 1." *Student Thinking*: 0 Because even though the response shows some concern for the student in providing him with enough time to think, it doesn't include any interpretation of what or whether the student understood. *Suggestions for Improvement*: 1 Because wait time is a general pedagogical suggestion and not based on the specific mathematics. *Depth of Interpretation*: 1 Because there is some interpretation around the idea of wait time idea.

Table 2. Average Percentage of Responses per Category Aggregated Across Rubrics and Clips by Topic.

Fractions				Variables, expressions, and equations				Ratio and proportions			
	0	1	2		0	1	2		0	1	2
Rubrics				*Rubrics*				*Rubrics*			
MC	45.76	43.30	10.94	MC	45.28	46.29	9.79	MC	39.98	48.59	11.43
ST	48.57	44.53	6.90	ST	60.61	33.09	6.29	ST	50.62	36.36	8.55
SI	67.99	22.83	9.18	SI	68.10	17.57	7.19	SI	66.42	17.01	12.28
DI	37.63	50.34	12.03	DI	51.75	36.64	11.60	DI	46.41	39.97	10.14
Clips				*Clips*				*Clips*			
1	50.56	40.37	9.07	1	56.46	36.59	6.95	1	50.84	38.88	10.29
2	50.58	37.74	11.68	2	59.51	34.27	6.22	2	62.32	30.74	6.94
3	55.42	37.65	6.93	3	60.75	32.00	7.25	3	72.12	44.11	8.77
4	51.21	35.08	13.71	4	53.41	35.73	10.85	4	53.81	33.69	12.50
5	44.61	44.31	11.08	5	64.90	29.57	5.53	5	55.98	34.93	9.09
6	56.43	35.13	8.44	6	55.80	34.90	9.30	6	51.79	39.71	8.49
7_a	38.69	37.93	23.38	7	60.91	31.86	7.23	7	54.93	32.45	12.62
7_b	48.89	38.33	12.78	8	61.89	28.31	9.80	8	42.77	39.62	17.61
8	46.97	46.39	6.64	9	52.11	35.56	12.32	9	60.14	28.26	11.59
9_a	54.86	40.35	4.79	10	54.98	33.89	11.14	10	54.21	36.06	9.74
9_b	51.83	43.46	4.70	11	56.85	36.54	6.61	11	31.73	57.69	10.58
10	49.35	45.37	5.28	12	51.84	37.62	10.54	12	56.34	36.60	7.06
11	50.43	41.13	8.44	13	54.53	34.07	11.40	13	54.31	33.01	12.68
				14	57.45	29.93	12.62				

Note. MC = mathematical content; ST = student thinking; SI = suggestions for improvement; DI = depth of interpretation.

Naïve Bayes Models

In this section, we present a brief introduction to naïve Bayes text classifiers in the context of our CVA teacher responses before providing additional details on some of the specifics of our analyses. For simplicity, we will only illustrate the workings of text classification for a single rubric; extending this simplified model to the four CVA rubrics is straightforward.

The input to our analysis was a collection of text responses, each paired with the score it had been given by a human rater on a particular rubric. As a first step, each teacher response was reduced to a compact representation in terms of a set of ''features,'' in our case, words that appeared in the response. Second, a machine learning algorithm was applied to inductively train a classifier based on associations between features (i.e., words) and scores.

In reducing our responses, we used a simple ''bag of words'' approach, in which all information about word order is discarded. This means that each text response was reduced to a list of the words that appeared somewhere in the response. Clearly, this is a dramatic simplification. Nonetheless, across a wide range of applications it

has been found that bag-of-words approaches give good results (Apte, Damerau, & Weiss, 1994; Dumais, Platt, Heckerman, & Sahami, 1998; Lewis, 1998; Sebastiani, 2002).

Once the text responses had been reduced, they were fed (along with their associated rubric scores) into the machine learning algorithm, which then produced a classifier model. As noted earlier, the particular algorithm we employed is called *naïve Bayes*. Naïve Bayes is simple enough that we can provide a nearly complete description here. To begin, note that our goal is to produce a classifier model that can be given a word list as input, and can produce a score as output. More specifically, in the case of naïve Bayes, we want our classifier to be able to calculate a probability for each score C (i.e., 0, 1, 2), given the list of words (w) contained in a response, and then select the score with the highest probability.

$$\mathrm{P}(C_i|w_1 \,\&\, w_2 \,\&\, w_3 \ldots).$$

Although we cannot directly extract probabilities of this sort from training data, we can estimate the overall probability for each score and each word—$\mathrm{P}(C_i)$ and $\mathrm{P}(w_j)$—simply by computing the fraction of responses in which they occur. Similarly, we can easily estimate the probability of each word, given a certain score, $\mathrm{P}(w_j \mid C_i)$, by counting the fraction of responses scored C_i which contain the word w_j. Finally, if we assume that the probability that a given word will appear in a response is independent of the other words that appear, we can use Bayes Theorem to compute the probabilities that are of interest:

$$\mathrm{P}(C_i|w_1, \,\&\, w_2, \,\&\, w_3 \ldots) = \frac{P(C_i)\prod_j P(f_j|C_i)}{\prod_k P(f_k)},$$

where the subscripts j and k both range over all the words in a response. Thus, to build a naïve Bayes classifier, we use a set of scored text data to estimate the probabilities $\mathrm{P}(C_i)$, $\mathrm{P}(w_j)$, and $\mathrm{P}(w_j \mid C_i)$, and then use these probabilities to determine the score for a new, unscored teacher response.[1]

In the approach outlined above, all the various probabilities are estimated from statistics gathered from a set of scored samples. In practice, one could manually adjust any of the values that are obtained. For example, the base probability of each code, $\mathrm{P}(C_i)$, could be adjusted based on human judgments about the relative probability of the code. We did not adjust values in this manner, in part to avoid problems of overfitting, an issue that is discussed below.

This approach is called ''naïve'' because it assumes that given a code, the probability that a word will appear in a response is independent of the other words that appear. Clearly, this assumption is false in our case. However, some investigators have reported reasonable success in applying naïve Bayes, even when the independence assumption is clearly false. When the amount of data to be analyzed is small, the limitations of naïve Bayes seem to be less problematic, and there is therefore less motivation to employ more complex methods (Kohavi, 1996; Lewis, 1998).

Our Analysis Procedure

The analysis begins with all the scored data for one rubric and one of the topics studied (e.g., DI for fractions). These data are then randomly divided into two parts: the data from 75% of participants is placed in a *training set*, and the data from the remaining 25% of participants is placed in a *test set*. Next, the responses in the training set are reduced to unordered word lists. These word lists, along with their associated scores, are then used to train a naïve Bayes classifier, as described above.

After the classifier is trained, data from the test set are then used to evaluate the performance of the classifier model. Each new response is converted to a word list and then input into the classifier model, which outputs a score for the response. These scores can then be compared with the scores given by human raters or evaluated in other ways. Because the results of this evaluation may vary based on how the data are divided into training and test sets, we used cross-validation techniques. For each topic we randomly drew 50 subsamples from our entire set of manually scored responses to train our classifier and averaged the results. Rerunning the analysis in this way has the additional benefit that it allows us to see whether the performance of the classifier is sensitive to particular features of the training and test sets. This, in turn, can help us understand how our results are likely to generalize to new data. The average cross-validated performance of the classifier based on the 50 runs may be taken as an estimate of the expected performance of a classifier trained with all our data, should that classifier be applied to new data collected in the future.[2]

In some applications, it is deemed prudent to include a *held-out* sample, a subset of the data that is reserved until an automated scorer is deemed complete and ready for publication. The advantage of employing a held-out sample is that it reduces problems of overfitting, in which a classifier is too narrowly tuned for the data at hand. For the present application, however, using a held-out sample seemed less feasible given the comparatively small samples that were available for developing our classifiers. As noted below, we have been careful to avoid the sort of excessive tuning that could lead to overfitting.

The description in the preceding paragraphs omitted a few critical details. First, when reducing each response to a word list, not all words are actually included. The final list is constructed as follows. To start, the algorithm looks across all responses for a given clip and compiles a complete list of all the words that appear in any of the responses for that clip—a *vocabulary*. Next, words that are very common but unlikely to differentiate responses in a meaningful way are put on a *stop list* and removed from the vocabulary. Our stop list included, for example, common words such as *and, the*, and *or*, and also some less common words that appeared in virtually every one of our responses (e.g., *teacher*). Finally, our vocabulary is further reduced to include only the 100 most frequently occurring words that remain.

Clearly, there is some art, and some trial and error, involved in the construction of the final word list. But we were also careful not to make choices that were too finely tuned to particularities of our data source, to avoid problems of overfitting. For example, we used the same stop list across all topics and clips and did not engage in extended tinkering so as to maximize the behavior of our classifiers with our unique data set.

A related issue is the specificity of our classifiers. Clearly, we must build a separate classifier model for each rubric. However, if it were possible to build one classifier model that could work across all topics and clips, there would be some important benefits. First, in the present work, we would have more data available for training, since we could pool the data across all clips. More profoundly, a general classifier could be applied to new clips and topics without additional training. However, a more general classifier is likely to be less accurate, since it cannot make use of features unique to individual clips and topics.

In the end, we settled on a hybrid approach, one that produces unique classifiers that are somewhat tuned for individual clips but that nonetheless capitalizes on the availability of additional data. To do this, we build a vocabulary by looking at all the responses to a particular clip. We then used this restricted vocabulary to extract feature sets from the entire set of responses associated with a particular topic area to train the classifier model. The result is that the classifier for a given clip can learn about which words tend to be associated with certain codes. But it does this without the distraction of words that do not tend to appear to responses to the particular clip.

The Mathematics Knowledge for Teaching (MKT) Instrument

As part of the original instrument development study, teachers had completed another measure of teacher knowledge, the MKT instrument in addition to the CVA assessments. The MKT instrument consists of multiple-choice items that measure different aspects of teacher knowledge that have been identified to be important for effective teaching through a job analysis approach.

For each of our three CVA assessments, we created a custom set of MKT items matched to the topic of our CVA scales. All three MKT scales consisted of 15 multiple-choice items, but some of the items in each scale were testlets, that is, a single stem was associated with two or more questions, resulting in a de facto larger number of items. Hence, the maximum score on the fraction MKT scale was 18, the maximum score on the RP scale was 27 and 35 on the VEE scale. All three scales were internally consistent as measured by Coefficient alpha (.77 for fractions, .84 for RPs, and .89 for VEE), which indicated that items measured a common dimension. In the original instrument development study, we had observed correlations of medium to large size between teachers' total and subscores on the CVA assessments and their scores on the respective topic matched MKT scales. In this study, we used the MKT scores to explore whether correlations based on computer-generated scores were similar in size and comparable to those observed for rater-assigned scores.

Results

We present results from this study in three sections. In the first section, we show in detail for one example response how scoring is determined using one of our naïve Bayes classifiers. In the second section, we present results detailing the relationship

between machine- and human-generated scores across the 50 randomly chosen subsets of manually scored reference texts. Here, we provide two different kinds of information. We report average correlations (Pearson product-moment and Spearman rank-order) between human- and computer-based total and subscores to evaluate the stability of teachers' relative standing across both scoring methods. Most of the time we are interested in these kinds of scale summary scores because we use them to either rank order teachers or to relate them to other scores or quantities of interest. We also report the weighted squared kappa statistic that represents the agreement between machine and human scores for ordinal data beyond chance agreement, aggregated across clips. In the third section, we report correlations between human- and computer-based total and subscores with another measure of teacher knowledge to explore whether machine scores produce correlations comparable in size to those observed for rater-assigned scores.

Looking Inside the Black Box of Machine Scoring

In Table 3, we show how one of our classifier models computes the best (most probable) scores for a single response to the clip described in Table 1, which is about creating equivalent fractions. The probability of each score is computed by beginning with the *base probability* of that score for a given rubric, which serves as an empirical prior, and then modifying that probability based on the specific words that appear in the given response. For example, Table 3 shows that the base probabilities for the mathematical content rubric for this clip are 46% for a score of 0, 42% for a score of 1, and 12% for a score of 2. This reflects the fact that given all reference set responses for this clip, scores of 0 were slightly more common than a score of 1, whereas a score of 2 was somewhat rare. In adjacent columns to the left in Table 3, we present the corresponding log likelihood values. Because log likelihood values can simply be added (rather than multiplied), it is easier to recognize the contribution of any particular value to the overall probability of a score.

After the classifier computes the probability of each score, it selects the score associated with the highest probability. For the example as shown in Table 3, a score of 1 is selected for the mathematical content rubric, since this score is associated with the highest probability (97%). A score of 1 is also most likely for student thinking (93%) and depth of interpretation (94%) whereas the response receives a score of 0 on the suggestions for improvement rubric (77%). For this example response, which is machine scored in perfect agreement with the rater-assigned scores (presented earlier), the probabilities associated with the final scores are very high, providing strong statistical evidence. However, one can imagine that for some responses two or even all three scores might have probabilities that are close, so that the selected final score is based on weaker probabilistic evidence, which might be one possible source of disagreements between machine and human scoring.

To better understand how the presence of specific words was contributing to each category score, we listed relevant words that were part of our example response

Table 3. Words, Associated Log Likelihoods, and Probabilities for Example Response 1.

"The teacher is *asking* the student *questions* to narrow down his search to *find equivalent fractions* I don't *know* if the boy *understood* the meaning of the *giant* I it was almost *like* the student was guessing and by judging how the teacher responded he would know if he was right or wrong."

Scoring Rubic	Mathematical content (−log)			Mathematical content (probabilities)		
Category scores	0	1	2	0	1	2
Base	−1.108	−1.251	−3.11	0.464	0.420	0.116
Final	−16.15	−0.044	−5.044	0.000	0.97	0.030
Asking	−4.347	−3.21	−3.257	0.049	0.108	0.105
Equivalent	−11.088	−3.044	−2.647	0.000	0.121	0.160
Find	−4.488	−3.308	−3.309	0.045	0.101	0.101
Fractions	−4.898	−1.546	−1.355	0.034	0.343	0.391
Giant	−9.503	−4.795	−4.232	0.001	0.036	0.053
Know	−5.307	−3.726	−2.861	0.025	0.076	0.138
Like	−3.327	−3.032	−2.432	0.010	0.122	0.185
Questions	−3.3	−2.962	−3.362	0.102	0.128	0.097
Understood	−4.898	−3.933	−3.733	0.034	0.066	0.075
Combined total of not-presented words	−6.032	−11.382	−16.889	0.015	0.000	0.000
Scoring Rubic	Student thinking (−log)			Student thinking (probabilities)		
Category scores	0	1	2	0	1	2
Base	−1.025	−1.207	−3.730	0.491	0.433	0.075
Final	−4.270	−0.109	−5.588	0.052	0.927	0.021
Scoring Rubic	Suggestions for improvement (−log)			Suggestions for improvement (probabilities)		
Code	0	1	2	0	1	2
Base	−0.525	−2.243	−3.415	0.695	0.211	0.094
Final	−0.38	−2.407	−4.536	0.768	0.189	0.043
Scoring Rubic	Depth of interpretation (−log)			Depth of interpretation (probabilities)		
Code	0	1	2	0	1	2
Base	−1.333	−1.068	−2.987	0.397	0.477	0.126
Final	−5.256	−0.095	−4.741	0.026	0.936	0.037

along with their probabilities and log likelihood values. Like the base probabilities, probabilities associated with individual words were obtained by analyzing the manually scored reference texts in the manner described above. As shown in Table 3, our example response contained the following relevant words: *asking, equivalent, find, fractions, giant, know, like, questions, understood*. To obtain the final probability for each MC score, the naïve Bayes classifier adds the respective log likelihood values for all relevant words that were part of the response text to the corresponding category base score. In addition, the classifier modifies the probability based on the words that do *not* appear in a response, since the absence of a word can also provide information as to the likelihood of a given response. For readability, we collapsed the contribution of all these not-present words to a single row in Table 3. Thus, the probability of an MC score of 0 is computed by summing $-1.108 + (-4.347) + (-11.088) + (-4.488) + (-4.898) + (-9.503) + (-5.307) + (-3.327) + (-3.3) + (-4.898) + (-6.\ 032) = -58.296$.

The respective log likelihood values for scores 1 and 2 were obtained in the same way, resulting in values of -42.189 for a score of 1 and -47.187 for a score of 2. These values, converted into probabilities yielded the final values of 0.00, 0.97, and 0.03 shown in Table 3 (after being normalized).

For this example, the computer-assigned scores matched the rater-assigned scores perfectly, but this was not the case for all responses. Next, we provide results on machine–human score agreement by rubric and topic.

Computer-Rater Agreement

Our first research question asks: *How similar are machine scores and human scores?* To answer this question, we aggregated the data at multiple levels, and we employed a number of measures. Our data set includes teacher responses to a total of 40 different video clips representing three different topic areas, each scored according to four rubrics.

Each CVA assessment uses multiple clips to assess a teacher's knowledge within a given topic area. Thus, the most important information is not how a teacher scored on individual clips; it is the aggregate score across the multiple clips that make up the assessment. Therefore, in comparing human and machine scoring, it was less important for us to determine whether human and machine scores matched for individual clips. We wanted to know if the computer's aggregate scores for a teacher, averaged across clips in an assessment, correlated with the aggregate human scores.

For the first set of results shown in Table 4, we display Pearson product-moment and Spearman rank-order correlations between machine and human raters both for total scores and subscores. These results are all computed by averaging across 50 cross-validation trials, as described above. We reasoned that if average correlations between rater and computer-generated scores exceed .80, an argument can be made for the convergent validity of machine scores with human scores. We also report averaged quadratic weighted Kappas as another measure of agreement between

Table 4. Average Correlations for Total and Rubric Subscores and Average Weighted Quadratic Kappas Aggregated Across Clips by Topic.

	Kappa quadratic				Pearson correlation				Spearman correlation			
	M	Min	Max	*SD*	*M*	Min	Max	*SD*	*M*	Min	Max	*SD*
Fractions												
Total	0.51	0.43	0.57	0.03	0.88	0.81	0.94	0.03	0.89	0.81	0.94	0.03
LI	0.56	0.46	0.65	0.04	0.83	0.72	0.94	0.05	0.83	0.67	0.94	0.06
MC	0.64	0.57	0.72	0.03	0.87	0.79	0.93	0.03	0.87	0.78	0.93	0.04
SI	0.37	0.26	0.50	0.05	0.65	0.43	0.80	0.09	0.64	0.43	0.80	0.10
ST	0.43	0.33	0.52	0.04	0.77	0.64	0.89	0.05	0.77	0.63	0.90	0.06
Ratio												
Total	0.51	0.41	0.58	0.04	0.86	0.79	0.94	0.03	0.86	0.77	0.95	0.04
LI	0.56	0.44	0.65	0.05	0.83	0.71	0.92	0.05	0.82	0.68	0.92	0.05
MC	0.64	0.55	0.70	0.03	0.90	0.85	0.93	0.02	0.90	0.85	0.94	0.02
SI	0.36	0.22	0.51	0.06	0.54	0.26	0.80	0.11	0.51	0.28	0.80	0.11
ST	0.47	0.33	0.55	0.04	0.81	0.67	0.90	0.04	0.81	0.70	0.88	0.05
VEE												
Total	0.55	0.48	0.62	0.03	0.89	0.81	0.97	0.03	0.91	0.83	0.96	0.03
LI	0.63	0.52	0.71	0.04	0.88	0.79	0.94	0.04	0.87	0.74	0.93	0.05
MC	0.62	0.55	0.69	0.03	0.86	0.80	0.92	0.03	0.86	0.76	0.93	0.04
SI	0.43	0.28	0.54	0.05	0.69	0.44	0.83	0.07	0.60	0.32	0.76	0.10
ST	0.46	0.36	0.55	0.05	0.82	0.67	0.90	0.05	0.83	0.74	0.91	0.04

Note. LI = Depth of Interpretation; MC = Mathematical Content; SI = Suggestions for Improvement; ST = Student Thinking; VEE = Variables, Expressions, Equations.

human and computer-generated scores. We computed these averages by pooling exact agreement information across clips within assessments. Again, the results are averaged across 50 cross-validation trials.

Table 4 shows three noteworthy results: Across all three topic areas, average correlations between human- and machine-generated total scores as well as for three of the four subscores (mathematical content, student thinking, and depth of interpretation) were around or greater than .80. (They ranged from.77 to .91.) This suggests that machine total scores and subscores measure a construct that is strongly related to the usable teaching knowledge construct measured by rater-assigned scores. Average correlations between computer-generated and rater-assigned scores for the suggestions for instructional improvement rubric were somewhat lower for all three assessments (between .49 and .69) with greater variation between different runs, indicating that performance of the classifiers depended to a larger degree on the particular training and testing subsamples. Maximum correlations were either approaching or at .80, whereas minimum correlations were as low as .19.

The reported quadratic weighted kappa values provide a measure of agreement beyond chance, weighting disagreements between nonadjacent categories more heavily than disagreements between adjacent categories. Similar to the correlations,

Table 5. Correlations Between Human Scores, Length of Response (Number of Words), and Machine Scores by Rubric and by Topic.

	Variables, expressions, and equations (VEE)		Ratios proportions (RP)		Fractions (F)	
Human scores	Total # words	Machine scores	Total # words	Machine scores	Total # words	Machine scores
Total score	.793**	.885**	.871**	.891**	.845**	.926**
LI score	.878**	.887**	.914**	.861**	.846**	.920**
MC score	.594**	.833**	.758**	.916**	.805**	.901**
SI score	.636**	.741**	.547**	.658**	.700**	.777**
ST score	.726**	.777**	.855**	.876**	.658**	.778**

Note. *N*(VEE) = 49; *N*(RP) = 52; *N*(F) = 45. LI = Depth of Interpretation; MC = Mathematical Content; SI = Suggestions for Improvement; ST = Student Thinking.
**$p < .01$.

Kappa values varied more across rubrics than across topics with consistently higher values for the mathematical content (.62 to .64) and the depth of interpretation (.56 to .63) rubrics and total scores (.51 to .55), indicating moderate to substantial agreement according to guidelines proposed by Landis and Koch (1977). Kappa values for the suggestions for improvement and student thinking rubrics (.36 to .43 and .43 to .47, respectively) were lower and indicated fair to moderate agreement. The Kappa statistic is sensitive to the prevalence of observed frequencies and some argue that it underestimates agreement for categories with higher frequencies, which makes interpretation of kappa values less clear.

Next, we compared correlations between scoring rubrics for each set of scores. Between-rubric correlations based on rater assigned scores ranged from .6 to .8 across topics, indicating that the four rubrics measured unique yet related aspects of usable teacher knowledge. The respective correlations based on machine scores were consistently higher across topics, ranging from .8 to .9, which suggests more redundancy, possibly a function of our efforts in the development phase to keep the classifiers more general.

Finally, in Table 5, we compare our human–machine correlations to the performance of a measure based solely on the number of words in a response. As noted elsewhere, length of response tends to be highly correlated with human scores (Attali, 2013). Overall, our classifiers do better than an analysis based solely on number of words, although improvements are variable and sometimes greater (e.g., for MC and SI) other times smaller (DI and ST).

We also computed partial correlations between computer- and rater-generated scores controlling for length of response to investigate whether the machine scores were related to the rater-assigned scores above and beyond response length. These correlations are shown in Table 6. Partial correlations remained moderate to large in size, indicating that a considerable part of the shared variance between human and

Table 6. Partial Correlations Between Human and Machine Scores After Controlling for Length of Response.

Human scores		Machine scores		
		Fractions	RP	VEE
Total score	Correlation	.722**	.503**	.648**
LI score	Correlation	.679**	.173 (p = .225)	.497**
MC score	Correlation	.699**	.794**	.791**
SI score	Correlation	.471**	.436**	.791**
	p value	.000	.000	.004
ST score	Correlation	.625	.489	.409

Note. df (F) = 42; df (RP) = 49; df (VEE) = 46. LI = Depth of Interpretation; MC = Mathematical Content; SI = Suggestions for Improvement; ST = Student Thinking; VEE = Variables, Expressions, and Equations.
**$p < .01$.

machine scores is independent of response lengths. Again, the only exception is the depth of interpretation rubric for the RP CVA, where the partial correlation is small and statistically not significant, indicating that length of response is confounded with the classifier's performance.

Overall, our results suggest some validity of the machine scores if rater-assigned scores represent the standard. Although the machine scores only reflect the occurrence of sets of words in teachers' responses, rank-ordering of teachers regardless of scoring method is fairly stable. Furthermore, our classifiers produced scores that measured more than simply length of response, while our results also raised some concerns about redundancy of scoring rubrics for machine scores.

There is some evidence supporting the generalizability of our algorithms across topic areas and potentially new CVA assessments. Results across all measures of agreement were fairly stable across the three CVA assessments, which might suggest a similar performance of scoring algorithms if additional CVA assessments that cover new content areas were to be developed, provided the clip selection process, the analysis prompt, and the rubrics, remain the same.

Internal Consistency of Human- and Machine-Generated Scores

To obtain some measure of reliability of our machine scores, we computed coefficient alpha as an indicator of internal consistency. Treating the video clips as items, we added rubric scores by clip to evaluate the degree to which our clips measured the same construct. For computer-generated scores internal consistency ranged from .90 to .95 depending on CVA assessment, suggesting that the clips measured a single construct or latent dimension. For rater-assigned scores, internal consistency was slightly lower, ranging from .89 to .93. Internal consistency estimates for the mathematical content, student thinking, suggestions for improvement, and depth of

interpretation rubric were .87, .84, .80, .81 for rater-assigned scores. Respective estimates for computer-generated scores were higher, possibly an artifact and a consequence of the higher between-rubric correlations we observed for machine scores. Values for the RP and VEE assessment were comparable.

Relating Computer- and Human-Based Scores to Another Measure of Teacher Knowledge

Next, we related computer-generated and human-assigned scores to another measure of teacher knowledge, the MKT instrument, to investigate the criterion-related validity of the machine scores. Table 7 reports correlations from a single scoring run.

Table 7 shows that, across all topic areas, correlations between computer-generated scores and the MKT were similar in size to those obtained for rater-assigned scores and the MKT, indicating that the relationships were stable regardless of scoring method. The results provide further evidence for convergent validity of the machine scores.

Discussion

In this study, we have explored the potential for automating the scoring of teachers' short, written responses to the CVA assessment, an innovative and promising assessment of teachers' usable teaching knowledge in mathematics. Teachers' scores on the assessments, when assigned by trained human raters, were reliable and for the topic of fractions predicted teachers' own teaching and their student learning. Machine scoring capabilities might ultimately make the assessments more practical to use and allow us to study the usefulness of the CVA approach more comprehensively.

As a first exploration of automated scoring, we constructed naïve Bayes text classifiers, using human-labeled responses as training data. We then set out to study the behavior of the resulting classifiers in multiple ways. Several interesting findings emerged from our study.

First, our results provided some evidence for the convergent validity of the machine scores with rater-assigned scores. Average correlations were around or above .80, for total scores and three of the subscores (mathematical content, student thinking, and depth of interpretation), for all three CVA assessments. We take this as indicating that the two scoring methods measure related constructs. Human–machine score agreement based on weighted quadratic Kappas was mostly moderate to substantial depending on the scoring rubric. We also found that our machine scores assessed more than length of response, which is known to be strongly related to human scores, adding to the evidence for the convergent validity of computer-generated scores.

Correspondence between machine and human scores for the suggestions for improvement rubric was somewhat lower. This might be due to the fact that the analysis prompt did not explicitly ask teachers to provide suggestions, resulting in fewer

Table 7. Correlations Between the MKT Teacher Knowledge Measure and Rater- and Computer-Generated Scores for the CVA Scales and Subscales by Topic.

		Classroom video analysis (CVA) assessment				
		MC	ST	SI	DI	Total
Fractions						
MKT	Rater-assigned scores	.39**	.34**	.45**	.48**	.49**
	Computer-generated scores	.44**	.48**	.47**	.50**	.49**
Ratio and proportions						
MKT	Rater-assigned scores	.63**	.55**	.45**	.56**	.63**
	Computer-generated scores	.59**	.60**	.53**	.59**	.59**
Variables, expressions, and equations						
MKT	Rater-assigned scores	.59**	.52**	.47**	.55**	.60**
	Computer-generated scores	.59**	.59**	.57**	.59**	.59**

Note. $n = 49$. LI = Depth of Interpretation; MC = Mathematical Content; SI = Suggestions for Improvement; ST = Student Thinking.
**$p < .01$.

overall suggestions, which left this rubric less well-defined. Although words such as ''should, would, could, might'' may serve as clear markers for making suggestions, they seem to have been too infrequent overall to be consistently included in the 100 most frequent word lists. In fact, for the fraction clip for which we illustrated automated scoring in detail for one authentic teacher response, only the word ''could'' was included among the 100 words included in the analysis. As noted by Streeter et al. (2011), task clarity is an important factor for machine scoring, just as it is for other types of assessments; in our case, revising the analysis prompt accordingly might increase the number of suggestions in teacher responses. Similarly, adding relevant indicator words to the list might improve our classifier performance.

We also examined the redundancy between rubrics for the two scoring methods. For human scoring, the between-rubric correlations ranged from .6 to .85, suggesting that the rubrics capture unique, yet related aspects of the usable teaching knowledge construct. In contrast, correlations for machine scores ranged from .8 to .94, indicating more redundancy. This might be related to the use of our ''hybrid'' approach, which leveraged the data across clips when creating the classifier for each clip. Future work needs to explore whether more clip specific classifiers might improve scoring accuracy of CVA responses and better preserve the uniqueness of each rubric.

Second, we found some evidence for the reliability of machine scores. Using clips as items, we computed a single score for each clip by adding the individual rubric scores and estimated internal consistency of the clips as a measure of reliability. We obtained coefficient alpha values ranging between .90 and .95 depending on topic, indicating that clips within each assessment measured the same construct. These values were similar to the values we obtained for rater-assigned scores (.90 to .93). A careful analysis of the actual score distributions needs to examine whether the high internal consistency of machine scores is an artifact of less overall variance for computer-generated scores across clips.

Third, correlations between computer-generated scores and scores from another measure of teacher knowledge, the MKT Instrument, were comparable in size to those based on rater-assigned scores. The results provide further evidence for the convergent validity of machine scores. Although this might not be surprising given the high correlations between machine and human total scores, it represents another piece of evidence supporting the use of the machine scores as indicators of usable teaching knowledge.

In this study, we focused on the properties of summary scores, which aggregated scores across clips, because these aggregate scores are typically used in assessment contexts to rank-order teachers, or to identify their particular strength and areas for improvement. However, we did observe some interesting variation across clips in the behavior of our classifiers. For example, for some clips, the scores given by our classifier more closely matched those given by human raters. There are a number of factors that might lead to this different performance across clips. The leveraging of data across multiple clips might have more adversely affected some clips than others. In

addition, clip quality (i.e., variation in stimulus strength across video clips), clip content, and the amount of measurement error in human scores for each clip, are other important factors. In the next phase of our work, we believe that a close analysis of this differing performance across clips will help us better understand how our classifiers function, and why they give the results they do. Ultimately, it may help us improve our current classifiers, and it might suggest that other classification approaches would be fruitful to explore.

At this point, it is too early to make any predictions about what roles automated scoring might play in the future design and deployment of the CVA assessments. It is possible that automated analysis might one day be used, in some circumstances, without any human scoring. Even if this is not the case, it seems likely that automated analysis can profitably be used as a complement to human analysis of the CVA, one that provides slightly different information. A larger question is whether our findings will translate to other kinds of more open-ended short answer items. We believe that the results reported here at least suggest that even a simple ''bag of words'' approach, can produce promising results, and thus suggest that further work in this area might be fruitful.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Work on this study was supported by the Institute of Education Sciences, Grant No. R305M060007.

Notes

1. This brief account omits some subtleties. A small number (in our case, 0.5) is added to the count of responses scored C_i, which contain the word w_j so that $P(w_j \mid C_i)$ is never zero. In addition, our computation includes the probabilities associated with words in the larger corpus that *do not* appear in a response.
2. Our use of repeated sampling validation rather than *k*-fold validation was dictated by the relatively small size of our corpus. For each topic area, we had responses from less than 250 teachers. We felt that a workable number of responses in the test set would be about 50. Thus, if we used *k*-fold validation we could only average 5 ''folds'' of the data.

References

Apte, C., Damerau, F., & Weiss, S. M. (1994). Automated learning of decision rules for text categorization. *ACM Transaction on Information Systems*, *12*, 233-251. doi: 10.1145/183422.183423

Attali, Y. (2011). *Automated subscores for TOEFL iBT® independent essays* (ETS Research Report No. RR-11-39). Princeton, NJ: ETS.

Attali, Y. (2013). Validity and reliability of automated essay scoring. In: M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181-199). New York, NY: Routledge.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, *37*, 573-595. doi:10.1137/1037127

Brew, C., & Leacock, C. (2013). Automated short answer scoring: Principles and prospects. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 136-153). New York, NY: Routledge.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*, 119-142. doi:10.3102/01623737025002119

Darling-Hammond, L., & Baratz-Snowden, J. (2007). *A good teacher in every classroom. Preparing the highly qualified teachers our children deserve*. Hoboken, NJ: Jossey-Bass.

Deerwester, S., Dumais, S. T., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391-407. doi:10.1002/(SICI)1097-4571

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). *Inductive learning algorithms and representations for text categorization* (pp. 148-155). Presented at the Seventh International Conference. New York, NY: ACM Press. doi:10.1145/288627.288651

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments*, *8*, 111-127. doi:10.1076/1049-4820(200008)8:2;1-B;FT111

Foltz, P. W., Laham, D., & Landauer, T. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, *1*(2). Retrieved from imej.wfu.edu/articles/1999/2/04/index.asp

Hill, H. C., & Ball, D. L. (2004). Learning mathematics for teaching: Results from California's Mathematics Professional Development Institutes. *Journal of Research in Mathematics Education*, *35*, 330-351.

Hill, H. C., Ball, D. L., Sleep, L., & Lewis, J. M. (2007). Assessing teachers' mathematical knowledge: What knowledge matters and what evidence counts? In F. Lester (Ed.), *Handbook for research on mathematics education* (2nd ed., pp. 111-155). Charlotte, NC: Information Age.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, *105*, 11-30. doi:10.1086/428763

Kersting, N. (2008). Using Video Clips as Item Prompts to Measure Teachers' Knowledge of Teaching Mathematics. *Educational and Psychological Measurement*, *68*(5), 845-861. doi:10.1177/0013164407313369.

Kersting, N. B., Givvin, K., Sotelo, F., & Stigler, J. W. (2010). Teacher's Analysis of Classroom Video Predicts Student Learning of Mathematics: Further Explorations of a Novel Measure of Teacher Knowledge. *Journal of Teacher Education*, *61*(1-2), 172-181. Doi: 10.1177/0022487109347875.

Kersting, N. B., Givvin, K. B., Thompson, B., Santagata, R. & Stigler, J. (2012). Developing Measures of Usable Knowledge: Teachers' Analyses of Mathematics Classroom Videos

Predict Teaching Quality and Student Learning. *American Educational Research Journal*, *49*(3), 568-590. Doi: 10.3102/0002831

Kohavi, R. (1996). *Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid*. Retrieved from http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/lecturas-clasificacion/NBTree.pdf

Landauer, T. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy and Practice*, *10*, 295-308. doi:10.1080/0969594032000148154

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159 -174. Doi:10.2307/2529310.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *Multiple values selected* (Vol. 1398, pp. 4-15). Berlin, Germany: Springer-Verlag. doi:10.1007/BFb0026666

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in NIPS*, *14*, 841-848.

Pauli, C. & Reusser, K. (2011). Expertise in Swiss mathematics instruction. In Y. Li & G. Kaiser (Eds.), Expertise in mathematics instruction: An international perspective. New York: Springer, 85-107. Doi: 10.1007/978-1-4419-7707-6_5.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47. doi:10.1145/505282.505283

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York, NY: Routledge.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. *International Encyclopedia of Education*, *4*, 20-26. doi:10.1016/B978-0-08-044894-7.00233-5

Shermis, M. D., & Hammer, B. (2012). *Contrasting state-of-the-art automated scoring of essays: Analysis*. Retrieved from http://dl.dropboxusercontent.com/u/44416236/NCME%202012%20Paper3_29_12.pdf

Streeter, L., Bernstein, J., Foltz, P., & DeLand, D. (2011). *Pearson's automated scoring of writing, speaking, and mathematics* (White Paper). Retrieved from http://kt.pearsonassessments.com/download/PearsonAutomatedScoring-WritingSpeakingMath-051911.pdf

Topol, B., Olson, J., & Roeber, E. (2010). *The cost of new higher quality assessments: A comprehensive analysis of the potential costs for future state assessments*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Zhang, H. (2004). The optimality of naive Bayes. In V. Barr & Z. Markov (Eds.), *Proceedings of the 17th International FLAIRS conference (FLAIRS2004)* (pp. 568-573). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.5794&rep=rep1&type=pdf

Zhang, M. (2013). *Contrasting automated and human scoring* (ETS Research Report No. RDC-21). Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf