

*Research Article***Validation of Automated Scoring of Science Assessments**

Ou Lydia Liu,¹ Joseph A. Rios,¹ Michael Heilman,¹ Libby Gerard,² and Marcia C. Linn²

¹*Educational Testing Service, Princeton, New Jersey*

²*Graduate School of Education, University of California, Berkeley, California*

Received 10 December 2014; Accepted 19 October 2015

Abstract: Constructed response items can both measure the coherence of student ideas and serve as reflective experiences to strengthen instruction. We report on new automated scoring technologies that can reduce the cost and complexity of scoring constructed-response items. This study explored the accuracy of c-rater-ML, an automated scoring engine developed by Educational Testing Service, for scoring eight science inquiry items that require students to use evidence to explain complex phenomena. Automated scoring showed satisfactory agreement with human scoring for all test takers as well as specific subgroups. These findings suggest that c-rater-ML offers a promising solution to scoring constructed-response science items and has the potential to increase the use of these items in both instruction and assessment. © 2015 Wiley Periodicals, Inc. *J Res Sci Teach* 53: 215–233, 2016

Keywords: automated scoring; c-rater-ML; science assessment

Constructed response items can both measure the coherence of student ideas and serve as reflective experiences to strengthen science instruction, yet they are rarely used in classrooms (Kuo & Wu, 2013). Although constructed response items can elicit complex thinking and reasoning in science learning, they are often avoided in assessments due to the cost of achieving objective scores (Lee, Liu, & Linn, 2011; Liu, Lee, & Linn, 2011). In some cases, the time required to score the items prevents teachers from providing timely guidance to students (Gibbs & Simpson, 2004) and may take time away from working with struggling students (National Council of Teachers of English, 2008). A recent review showed that due to challenges in scoring, assessments that require extended responses are rarely used in classrooms (Kuo & Wu, 2013). Automated scoring of constructed response items has the potential to mitigate the scoring challenges and increase use of these items in science education (Linn et al., 2014; Liu et al., 2014).

This study investigates an automated scoring tool recently developed by Educational Testing Service (ETS) called c-rater-ML. The c-rater-ML system uses a machine learning technique called support vector regression to model the relationship between students' responses and scores (see, e.g., Smola & Schölkopf, 2004). We investigate three questions: (a) how accurate is c-rater-ML in scoring constructed-response science assessments with complex scoring rubrics? (b) how well does c-rater-ML assess performance differences for subgroups (based on gender, home

Contract grant sponsor: National Science Foundation; Contract grant number: 1119670.

Correspondence to: Ou Lydia Liu; E-mail: lliu@ets.org

DOI 10.1002/tea.21299

Published online in Wiley Online Library (wileyonlinelibrary.com).

language, computer use)? and (c) what accounts for large scoring discrepancies when performance differences occur?

Background

Over the last two decades, automated scoring has been widely developed and used in a variety of content domains, such as mathematics (Bennett & Sebrechts, 1996; Sandene, Horkay, Bennett, Braswell, & Oranje, 2005), science (Linn et al., 2014; Liu et al., 2014; Nehm, Ha, & Mayfield, 2011), and language testing (Bernstein, Van Moere, & Cheng, 2010; Higgins, Zechner, Xi, & Williamson, 2011), to name a few. Furthermore, in assessing written responses across content domains, automated scoring has been used to evaluate rubric dimensions, such as content (Attali & Powers, 2008; Dzikovska et al., 2013; Leacock & Chodorow, 2003; Mitchell, Russell, Broomhead, & Aldridge, 2002; Nielsen, Ward, & Martin, 2008; Sukkarieh & Bolge, 2010) and quality (Burstein & Marcu, 2002; Foltz, Laham, & Landauer, 1999). The accuracy of automated scores depends on a number of factors, including the content domain, the complexity of the tasks, the levels of the scoring rubrics, and the number of responses available to build the automated scoring models.

In this section, we focus on the review of prior work conducted in the domain of automated scoring of science content. Before the review, we describe the criteria used to evaluate the accuracy of automated scoring so readers have the background knowledge to understand the discussion in the literature review.

Evaluation Criteria

To evaluate the accuracy of automated scoring, prior research typically uses *Quadratic-weighted kappa*, *Pearson product-moment correlation*, *degradation from the human/human score agreement*, *standardized mean score difference*, and the F_1 statistic (Dzikovska et al., 2010; Williamson, Xi, & Breyer, 2012). In this study we use quadratic-weighted kappa and Pearson correlation, the most common methods for evaluating the accuracy of automated scoring. The kappa coefficient indicates the percentage of score agreement between two raters beyond what is expected by chance and has valid values ranging from -1 to 1 , with -1 indicating poorer than chance agreement, 0 indicating pure chance agreement, and 1 indicating perfect agreement (Fleiss & Cohen, 1973). When the categories are ordinal, it is useful to assess the degree of disagreement. This can be done by applying quadratic weights, which are proportional to the square of the number of categories that differ between two ratings, to kappa so that large disagreements in scores are more heavily penalized. This approach is referred to as quadratic-weighted kappa (K_{QW}). Landis and Koch (1977) proposed the following as standards for the strength of agreement for the kappa coefficient: poor (≤ 0.00), slight ($0.00-0.20$), fair ($0.21-0.40$), moderate ($0.41-0.60$), good ($0.61-0.80$), and very good ($0.81-1$). The Landis and Koch rule is applied to evaluate the accuracy of automated scoring in this study.

The Pearson correlation has also been used as a common criterion to evaluate the consistency between human and machine scores. We followed Cohen's (1968) rules for describing the magnitude of Pearson correlations: none ($0-0.09$), small ($0.10-0.30$), moderate ($0.31-0.50$), and large ($0.51-1.00$).

Clinical interviews have also been proposed to triangulate the accuracy of automated scores (Beggrow, Ha, Nehm, Pear, & Boone, 2014). Beggrow et al. (2014) compared human and automated scores of college students' written explanations of the causes of evolutionary change to cognitive interviews with students. They found that automated scores of written explanation showed strong correspondence to students' oral interview measures and that the automated scores were able to capture students' both scientific and naïve ideas as accurately as human scores.

However, given the extensive work associated with cognitive interviews, this approach hasn't been used widely in the evaluation of automated scoring.

Automated Scoring of Science Content

Mohler, Bunescu, and Mihalcea (2011) described and evaluated different variations of an automated scoring system on a dataset of introductory computer science questions. Their discussion focused on specific natural language processing and machine learning methods. However, the fairly high levels of human-machine agreement that they reported suggest that automated scoring could be useful pedagogically for the domain they studied, though room for improvement remains. For example, the highest level of human-machine agreement that they reported in Pearson correlation was 0.52, which is close to but not quite as high as, the reported human-human agreement of 0.59.

Ha, Nehm, Urban-Lurain, and Merrill (2011) tested machine scoring of college students' written explanations of evolutionary change. They used a concept-based scoring tool designed by researchers at Carnegie Mellon University (Mayfield & Rosé, 2010), called the Summarization Integrated Development Environment (SIDE), which was similar to c-rater in that it required the identification of key concepts from student responses for automated scoring. In this study, two human raters identified the presence or absence of five key concepts related to evolution. The authors reported that in terms of identifying the key concepts, SIDE achieved satisfactory agreement with human raters, with kappa values larger than 0.80. This study also investigated the impact of sample size on SIDE's scoring accuracy and noted that the automated scoring models trained on a larger corpus ($n \approx 1,000$) did not necessarily yield better results than models trained on a small corpus ($n \approx 500$).

The same SIDE program was used by Nehm et al. (2011) to score undergraduate students' explanation of evolution. SIDE showed consistent agreement with human raters (i.e., $\kappa > 0.80$). The authors also investigated the impact of response length (i.e., short or long responses) on scoring accuracy and found that response length did not significantly affect scoring performance indicated by kappa and percent agreement. Findings from this study confirmed that automated scoring could be a cost-effective way for assessing student complex science knowledge.

Other researchers have used automated tools for scoring in science education. For example, Haudek, Prevost, Moscarella, Merrill, and Urban-Lurain (2012) used SPSS Text Analysis when analyzing college students' explanations of acid-base behavior of biological functional groups. Through lexical analysis, students' responses were classified into 27 categories, where a category represented a cluster of similar terms defined by either the scoring tool and/or the user. Subsequent validation revealed that key lexical categories were able to predict expert ratings with satisfactory reliability. The authors concluded that computerized lexical analysis has potential in extracting common features from a large number of student responses and that such functionality may help instructors gain class-level insights of student thinking. The same SPSS tool was also used by Weston, Parker, and Urban-Lurain (2013) to score undergraduate students' responses to a question on cell metabolism. SPSS Text Analysis showed good agreement with human raters in classifying responses using an analytic rubric designed by researchers. The authors noted that it took two human raters over 50 minutes to score 50 sample responses, while it took the machine only 15 minutes to score 360 responses, clearly demonstrating the efficiency of automated scoring as an advantage over human scoring.

Furthermore, in the field of computational linguistics, automated scoring for science tutoring and assessment has been a topic of recent interest. For example, Dzikovska et al. (2013) described an evaluation of automated scoring systems from nine research groups on two science datasets, one consisting of scored responses to questions in a tutoring system for electricity and electronics

(Dzikovska et al., 2010) and the other consisting of scored responses to questions in 15 different science domains (Nielsen, Ward, Martin, & Palmer, 2008). They found that the best systems were able to provide performance deemed adequate for some applications. For example, they reported an average F_1 score above 0.80 for multiple systems for the task of classifying responses as correct versus incorrect using the dataset from Dzikovska et al. (2010). Note that F_1 is a statistic typically used to evaluate the accuracy of automated scoring and ranges from 0 to 1, with the best value being 1. They also observed lower performance (e.g., F_1 scores around 0.70 or lower) for certain situations, such as when item-specific training data is unavailable.

Liu et al. (2014) reported using c-rater, a concept-based automated scoring engine developed by ETS, in scoring inquiry science items with complex scoring rubrics. For c-rater scoring, one or more model responses needs to be identified and the linguistic features of the model responses are analyzed using natural language processing techniques. Such linguistic features are then applied to evaluate students' responses to determine the presence or absence of key concepts (Sukkarieh & Blackmore, 2009). A c-rater analytic scoring rubric combines the concepts following specified scoring rules to compute a score (e.g., concept one and two equals score 3, concept three equals score 2, etc). After the analytic scoring rubric is finalized, two human raters score the responses against the rubric, and their scores are then compared to the later c-rater scoring.

In a pilot study, Liu et al. (2014) scored four science items using c-rater and found that the items they tested showed moderate to good agreement between automated and human scores ($K_{QW} = 0.46$ to $K_{QW} = 0.64$). They also identified a few challenges in using the concept-based scoring method. First, it is time consuming to identify all key concepts in model responses. Sometimes the number of key concepts is unmanageable for human raters. Figure 1 shows the platform where the human coding takes place. This item states "A metal spoon, a wooden spoon, and a plastic spoon are placed in hot water. After 15 seconds which spoon will feel hottest? Explain your choice." Each rater is required to rate this single response against each of the key concepts (denoted as C1, C2, and so on) identified. Note that there were close to 10 concepts for this item but only one was shown for illustration purposes). Specifically, for each concept, the rater needs to rate whether the concept is A (absent), P (present), or N (not applicable). And they also need to put in a note where confusion or disagreement might occur with regard to the absence or presence of a concept. If there are 1,000 responses for this item with 10 concepts, then a single rater needs to rate the responses 10,000 times (multiplying the number of responses and number of concepts). To ensure the accuracy and consistency of the scoring, another rater is also required to conduct the

Rater: dkirkpatrickScore Name: SpoonBatch: 17785Result/Response ID: 183927/236053950 of 50Logout

Item Text:

A metal spoon, a wooden spoon, and a plastic spoon are placed in hot water. After 15 seconds which spoon will feel hottest? Explain your choice.

Hide Item Text

Spoon

The metal spoon EXPLANATION: I chose the metal spoon because metals are very good conductors of heat.

Concepts

C1 : The Metal Spoon Feels like a Different Temperature than its Actual Temperature
The metal spoon will feel the hottest, but it will still be the same temperature as all of the other spoons OR The metal spoon feels hotter than it actually is OR The metal spoon feels like a different temperature than its actual temperature
► Note:

A P N

☐ A ☒ P ☐ N

☐ A ☐ P ☐ N

Quotes / Highlights

Submit

Figure 1. c-rater scoring platform.

Journal of Research in Science Teaching

scoring at the same scale. The cumbersome nature of the scoring may have negatively impacted the scoring accuracy of the human raters, which serves as a key foundation for the validation of automated scoring.

Second, it is challenging, if not impossible, to exhaustively capture all the key concepts, especially for invalid science ideas which can be idiosyncratic and cannot always be categorized based on certain common ground. If each unique invalid science idea is included as a concept, then there will be too many concepts in the analytic rubric. Third, the relationship between the analytic rubric and the original human scoring rubric is not fully specified. When the original rubric is holistic, and the c-rater analytic rubric is concept based, the analytic rubric may not faithfully represent the original rubric in terms of construct coverage. As a result, the authors concluded that although automated scoring has shown potential to alleviate teachers' time spent on scoring constructed-response items, for its current design, it cannot replace humans in scoring science content as it requires too much fine tuning.

Linn et al. (2014) investigated the application of automated scoring in facilitating immediate feedback when teaching complex science topics such as photosynthesis or mitosis for middle school students. Students were randomly assigned to two feedback conditions: the immediate guidance condition assigned based on c-rater scoring, and the delayed guidance condition assigned by the teacher. Also using c-rater, the authors found that automated feedback was as effective as teacher feedback in terms of prompting students to review instruction and revise responses, and the score gains students achieved were comparable in both conditions. This study sheds new light on the possible applications of automated scoring in classroom level assessments.

Moharreri, Ha, and Nehm (2014) used EvoGrader, an online, open-source formative automated scoring tool, to score college students' responses of evolutionary explanations. The EvoGrader database includes automated scoring models for 86 different items related to natural selection. Undergraduate biology instructors can upload a response file containing unlimited numbers of evolutionary explanations to any of these items and obtain automated scores. Over 2,000 responses were scored by EvoGrader, and the automated scores were comparable to human scores in terms of identifying scientific and naïve concepts.

Table 1 presents more details from some of the studies reviewed above when they are available from the studies. The research to date reveals the potential of automated scoring in rating science content. However, given the many available tools designed for automated scoring, questions arise as to how to select a desirable automated scoring tool that can be easily implemented without requiring too much effort from human raters. Applications of automated scoring in classrooms largely hinge on the usability of such tools. Many of the above-reviewed automated scoring tools require a significant amount of human effort and expertise in implementation. For example, the SPSS Text Analysis requires human experts to build term libraries and identify text extraction rules in order for the text analyses to take place (Nehm et al., 2011; Nehm & Haertig, 2012). Compared to the exploratory nature of the SPSS Text Analysis, SIDE is considered an improvement as a confirmatory text analysis tool. However, in the experiments described by Nehm et al. (2011) and Nehm and Haertig (2012), for validation with human scores SIDE still requires human identification and assignment of key concepts, similar to c-rater. In addition, not all items can be decoded into a few meaningful concepts. In many cases, the knowledge underlying the items is so complex that it requires the presence of a large number of concepts in the scoring rubric, which will likely prevent human raters from producing reliable annotations across concepts (Liu et al., 2014). Therefore, scoring tools that can achieve good human-machine agreement but require substantial manual effort for tuning to new assessment items may have limited applications.

Table 1
Previous studies evaluating the accuracy of content-based automated scoring

Author & Year	Automated Scoring Tool	Population	Sample Size	Domain/ Topic	Scoring Level	Response Length	<i>M</i>	<i>SD</i>
							Evaluation Criterion	
Dzikovska et al. (2010)	BEETLE II	College	73	Physics	5	1–2 sentences	0.69 ^a	—
Dzikovska et al. (2013)	BEETLE II	College	3,000	Electricity & electronics	2,3,5	1–2 sentences	~0.70 ^b	—
Dzikovska et al. (2013)	BEETLE II	College	10,000	15 Science domains	2,3,5	1–2 sentences	~0.70 ^b	—
Ha et al. (2011)	SIDE	College	1,050	Biology	4	Average 45 words	0.80 ^a	—
Haudek et al. (2012)	SPSS text analysis	College	1,172	Science	3	Average 37 words	0.90 ^c	—
Liu et al. (2014)	c-rater	Middle school	321–412	Science	5	1–3 sentences	0.55 ^a	0.08
Liu et al. (2014)	c-rater	Middle school	321–412	Science	5	1–3 sentences	0.63 ^c	0.04
Moharreri et al. (2014)	EvoGrader	College	2,200	Biology	4	—	0.88 ^a	0.07
Mohler et al. (2011)	Graph alignment	College	2,273	Computer science	6	—	0.52 ^c	—
Nehm et al. (2011)	SIDE	College	565	Biology	4	—	>0.80 ^a	0.02
Nehm and Haertig (2012)	SPSS text analysis	College	330	Natural selection	7 concepts	—	0.60–1.00 ^a	—

Note: ^aKappa.
^bF1.
^cPearson correlation.

Group Differences

Previous research involving assessment items similar to the ones used in this study suggest little gender difference and some advantage for students who speak English as a first language as well as students who use a computer for homework (Liu, Lee, Hoftstetter, & Linn, 2008, 2011; Liu, Ryoo, Linn, Sato, & Svihla, 2015). As part of the validation of automated scoring, it would be important to examine if automated scores show similar patterns of group differences. Although automated scoring has been extensively researched, the focus on group differences is rarely seen in empirical studies. A few studies that deal with this topic were conducted by researchers in the validation of *e-rater*[®], an automated scoring tool developed by ETS for scoring writing quality (Bridgeman, Trapani, & Attali, 2012; Burstein & Chodorow, 1999; Chodorow & Burstein, 2004). Burstein and Chodorow (1999) reported that Arabic and Spanish speakers tend to benefit from human scoring while Chinese speakers tend to benefit from machine scoring. Chodorow and Burstein (2004) conducted a follow up study and found an interaction between scoring method (e.g., human or automated scoring) and language on one of the seven prompts tested. On this

particular prompt, Arabic and Japanese speakers had higher *e-rater* scores while Spanish speakers received equal scores from human rater and *e-rater*. Bridgeman et al. (2012) compared human and *e-rater* scores by gender, ethnicity, and country on the Graduate Record Examinations (GRE) and the Test of English as a Foreign Language (TOEFL). On both tests, Chinese test takers received higher scores from *e-rater* than from human raters. On TOEFL, Arabic and Hindi speakers received lower scores from *e-rater* than from human raters. Chinese test takers' advantage with *e-rater* was speculated to be related to the rigorous test preparation in mainland China, which involves memorizing large chunks of text that can possibly be included in essays. While it is straightforward for human raters to evaluate the relevance of the text to the required topic, it is challenging for machine scoring to tell when the text is off topic and therefore assigns a high score as long as the essay is grammatically correct and well-structured.

Although the above-reviewed studies focus on writing quality rather than science content, their findings point to the importance of difference investigations. Unintended differential performance by subgroup can occur as a result of the switch from human scoring to automated scoring.

Methods and Analyses

c-rater-ML

Given the known challenges of using *c-rater* and other similar programs in prior investigations (Ha et al., 2011; Liu et al., 2014; Nehm et al., 2011), another automated scoring tool, *c-rater-ML* was used in this study¹.

As mentioned earlier, the *c-rater-ML* system uses support vector regression to model the relationship between students' responses and scores (see, e.g., Smola & Schölkopf, 2004) and to identify a mapping from text responses to scores. Support vector regression is fundamentally similar to multiple regression in that it models relationships between predictor variables and independent variables, but it is specially designed to handle large datasets with large numbers of correlated predictors. To perform support vector regression, the *c-rater-ML* system uses an open-source software package called SciKit-Learn Laboratory (<http://github.com/EducationalTestingService/skll>). The model produced by support vector regression can be used to automatically predict scores for new student responses which were not included in the data used to estimate the model. Note that the use of machine learning techniques such as support vector machines for automated content scoring is not novel. For example, the system described by Mohler, Bunescu, and Mihalcea (2011) also used support vector machines to predict scores from linguistic features of student responses, as did many of the systems developed for the Automated Student Assessment Prize competition for short answer scoring (<http://www.kaggle.com/c/asap-sas>).

The *c-rater-ML* approach addresses the challenges of other scoring tools described above because it learns a statistical model from a set of previously scored responses rather than requiring descriptions of the key concepts that are relevant to the assessment item, as in the *c-rater* approach. The level of effort to estimate a *c-rater-ML* model based on a set of scored responses is much less than the level of manual effort required to describe key concepts. The design of *c-rater-ML* represents the state-of-the-art techniques in computational linguistics and related fields (e.g., text categorization for information retrieval). It has been informed by considerable research by ETS on automated content scoring, including work on prototype systems that have demonstrated excellent performance in public competitions and shared tasks (e.g., the Automated Scoring Assessment Prize [ASAP] in 2012 sponsored by the Hewlett Foundation, the Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge in 2013; Heilman & Madnani, 2013). The

choice of c-rater-ML in this study is based on two primary considerations: (a) it demonstrated satisfactory human-machine agreement, and (b) it is relatively straightforward to apply to new items.

For each new item, about two-thirds of the responses scored by human raters are used to estimate the parameters in the c-rater-ML automated scoring model, and the other third are used for the blank evaluation. Typically, around 500–1,000 responses are needed to reliably build the automated scoring model for an item.

When estimating a model, the c-rater-ML system computes a large set of predictor variables (or “features,” to use natural language processing terminology) from human-scored student responses used to train the model. These predictors consist of a variety of linguistic features that are commonly used in modern natural language processing (NLP) systems. For example, an item states:

“Burning coal to produce electricity has increased the carbon dioxide content of the atmosphere. What possible effect could the increased amount of carbon dioxide have on our planet?”

- A warmer climate.
- A cooler climate.
- Lower relative humidity.
- More ozone in the atmosphere

Explain your answer.”

Some sample responses include “*Carbon dioxide is a greenhouse gas. It traps the infrared radiation and creates heat*” and “*Burning coal to produce electricity has increased the amount of carbon dioxide in the atmosphere. Effect could the increased amount of carbon dioxide.*” When scoring students’ explanation of the possible effect that the increased amount of carbon dioxide may have on our planet, c-rater-ML evaluates sequences of words in the response, such as “carbon dioxide.” It also examines sequences of characters in the response which may help model variations in spelling or morphology. For instance, for the word “dioxide,” possible characters include “diox,” “ioxi,” “oxid,” etc. In addition, the automated scoring tool will analyze syntactic dependencies of the responses, among other features. For example, “coal” is the object of “burned.” Furthermore, it evaluates features based on semantic roles (e.g., “coal” is the second argument of the predicate “burned”) and response length bin features (e.g., whether the response was longer than eight words). The strongest predictors are typically the sequences of characters, and the other features, such as the syntactic dependencies, provide additional coverage of relevant linguistic constructs.

Eight science explanation items were scored using c-rater-ML. The following criteria were applied in item selection: (a) they had been previously administered in a science assessment to middle school students; (b) at least 500 total student responses were available on those items (note that two-thirds of the responses are used for model building and one third for blank evaluation; the sample sizes reported in Table 4 are for blank evaluation); and (c) the items could be incorporated into an inquiry science unit to provide automated guidance on central concepts. Instruction units incorporating the c-rater-ML items with automated guidance are in Table 2.

As a first step in c-rater-ML scoring, all responses were scored using the knowledge integration rubrics by humans with experience and expertise in domain and knowledge integration scoring. The knowledge integration scoring rubrics (Linn & Eylon, 2006) reward students’ ability to use evidence to make connections among ideas. The items were scored on a five-point scale (see Table 3) indicating progressively more links made among normative ideas about the targeted

Table 2
c-rater-ML items, learning goal, inquiry units

Item	Assessment Used to Gather Model Data	Grades in Which Model Data Collected	Dates of Model Data Collection	Learning Goal	Inquiry Unit for Automated Guidance
1	Cumulative end of year test	6–8th	2009–2014	Energy transformation	Photosynthesis
2				Energy transfer, transformation	Photosynthesis
3				Energy transfer	Thermodynamics
4				Convection, density	Plate tectonics
5	Unit pre and post test;	7th	2012–2014	Cell division phases, mechanisms, and structures	Mitosis
	Cumulative end of year test	6–8th	2005–2008		
6	Benchmark and delayed posttest	6–8th	2005–2008	Plate movement mechanisms	Plate tectonics
7	Benchmark and delayed posttest	6–8th	2005–2008	Dominant and recessive allele transmission	Simple inheritance
8	Unit pre and post test;	8th	2012–2014	Position/time graphs	Graphing stories
	Benchmark and delayed posttest		2008		

science topic (i.e., irrelevant answers, incorrect understanding, partially correct understanding, fully correct understanding, and advanced understanding). Prior research shows that these items have good psychometric properties including high reliability, lack of differential item functioning for subgroups, and satisfactory item fit statistics (Liu et al., 2008). Studies also show that questions designed to measure knowledge integration validly assess students' conceptual understanding (Linn & Eylon, 2006; Liu et al., 2011). For model evaluation, only responses that were scored by both human raters and the automated algorithm were included, which across the eight items ranged from 379 to 1,922.

Each of the eight items was coded by two humans who had domain expertise and prior coding experience using the knowledge integration scoring rubric. For each item, the two humans individually coded 50 student responses that spanned a range of possible scores. Coders then compared their two scores and identified disagreements. Coders discussed disagreements until reaching consensus. When the two coders could not agree on the scoring for a response, they introduced a third rater until consensus was reached. The process was iterated until the human–human agreement reached 90% on average. Once consensus was reached, the coders clarified the

Table 3
General knowledge and integration scoring rubric

Score	Description
1	Off task; Answers such as “I don’t know”
2	Incomplete: Non-normative ideas; Irrelevant ideas; Repeats the question
3	Partial: Isolated normative idea; Normative and Non-normative ideas
4	Full: A connection between two valid, key ideas
5	Complex: More than two connections between valid scientific ideas

rubric, split the remaining responses, and scored them individually. Coders selected a random sample of 20 responses intermittently for cross-scoring to ensure consistency. A limitation for this dataset was that no inter-rater reliability was recorded for each individual item except that the rule of 90% agreement for the initial scoring was applied to all items.

The second step of c-rater-ML scoring was to build automated models based on student responses. Specifically, the responses for each item were randomly split into two sets. The first, consisting of approximately two-thirds of the data, was used for training a model and for preliminary evaluations. The second, consisting of the remaining one third, was held out for the final testing, or validation, of the model estimated from the training set. Quadratic weighted kappa and Pearson correlations were used to evaluate the agreement between c-rater-ML and human scores on the held-out testing set.

As previous research has demonstrated subgroup differences across human and automated ratings (Bridgeman et al., 2012), we compared the equivalence in scoring procedures by gender, language, and use of computer for homework. *t*-tests were conducted to compare the group differences and an effect size in standardized mean difference was provided for each of the comparisons (Cohen, 1988). The classification of the effect size is as follows: negligible difference if $d \leq 0.20$, small if $.21 \leq d \leq 0.50$, moderate if $0.51 \leq d \leq 0.80$, and large if $d \geq 0.81$. Lastly, scoring analysis was conducted for the items showing large scoring discrepancies between automated and human scores.

Results

Agreement Between Human and c-rater-ML Scores

Overall, we found good agreement between c-rater-ML and human scores that held up for subgroups. We analyze the human and machine scores for discrepancies. Table 4 shows the consistency between human and automated scores as indicated by kappa and correlation. According to the Landis and Koch (1977) rule, two of the eight items showed very good agreement (i.e., κ over 0.80), while the remaining six items showed good agreement (i.e., κ between 0.60 and 0.80) between human and automated scores. In terms of Pearson correlation, all of the items showed large correlations (i.e., larger than 0.50) between human and automated scores, according to Cohen’s (1968) guidelines.

Table 4
Consistency among human and automated scores

Item	<i>n</i> ^a	Human		Automated		Agreement	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>K</i> _{QW}	<i>r</i>
1	1,922	2.33	0.79	2.32	0.78	0.90	0.91
2	1,324	2.63	1.09	2.59	1.00	0.86	0.87
3	242	2.43	0.85	2.45	0.75	0.72	0.71
4	404	2.80	0.91	2.84	0.82	0.73	0.79
5	377	2.39	0.71	2.39	0.62	0.62	0.66
6	1,284	2.64	0.88	2.64	0.82	0.80	0.83
7	712	2.78	0.99	2.78	1.00	0.79	0.82
8	1,045	3.25	0.76	3.19	0.72	0.76	0.79

Note: ^aThe sample sizes reported here are the responses used for blank evaluation, which are about 1/3 of the total responses. The rest of the 2/3 was used for model building.

Subgroup Differences

Gender. Using automated scoring, there were no significant differences for gender across all items. Although human scores showed that three items (Items 1, 3, and 8) favored (i.e., provided higher scores for) males and one item (item 7) favored females, the effect size differences were negligible for all items, except for the item favoring females ($d = 0.26$). Overall, the effect sizes for standardized mean difference between males and females were low (ranging from -0.22 to 0.10 for automated scoring and from -0.19 to 0.26 for human scoring; Fig. 2), suggesting no consistent differences between human and machine scores.

Language. Using both automated and human scoring, only item 1 showed a statistically significant difference between students who speak English as their first language and those who do not. The magnitude of differences across scoring procedures was negligible. Both automated and human scoring classified six of the eight items as possessing negligible mean differences (items 1, 2, 4, 6, 7, and 8), while items 3 and 5 were classified differently across scoring procedures. More specifically, item 3 was classified as showing a small difference for automated scoring ($d = 0.24$) and a negligible difference ($d = 0.04$) for human scoring. The opposite trend was observed for item 5 with classification for automated scoring being negligible ($d = 0.03$) but small for human scoring ($d = 0.36$; Fig. 3).

Computer Use. Using automated scoring, two items (items 3 and 4) showed statistical significance in favor of students who use a computer for homework, while four items (items 1, 2, 3, and 4) showed statistical significance favoring computer users when using human ratings. In terms of effect sizes, the automated and human scoring procedures agreed on three items that showed negligible differences (items 1, 2, and 5) and two items (items 4 and 7) that showed small differences. In addition, both scoring procedures agreed that item 3 possessed large mean differences across computer use groups, however the extremely small sample sizes ($n = 4$ to $n = 52$) for non-computer users suggest that these results need replication. Across all items, only item 6 differed in magnitude classification by scoring procedure (Fig. 4).

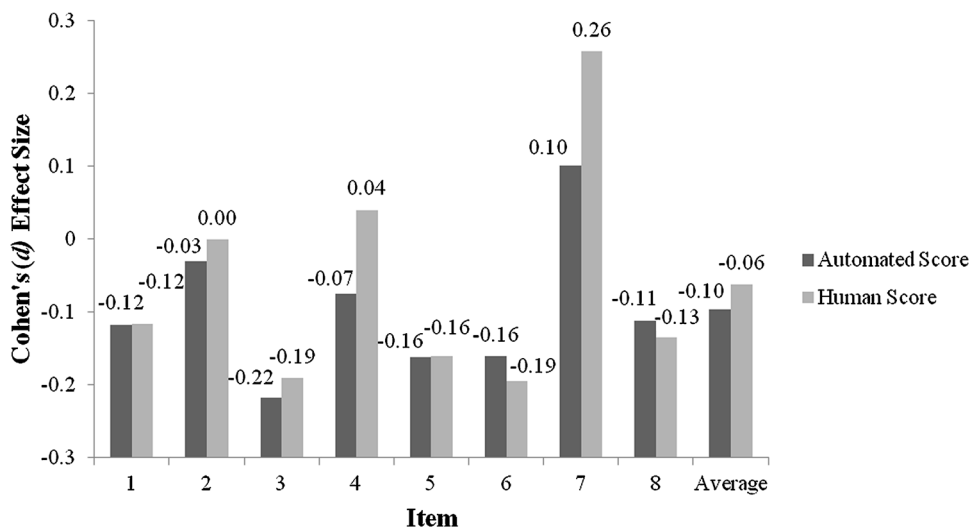


Figure 2. Effect size differences by gender across scoring methods and items. The direction of the effect size results were based on subtracting the mean of females from the mean of males.

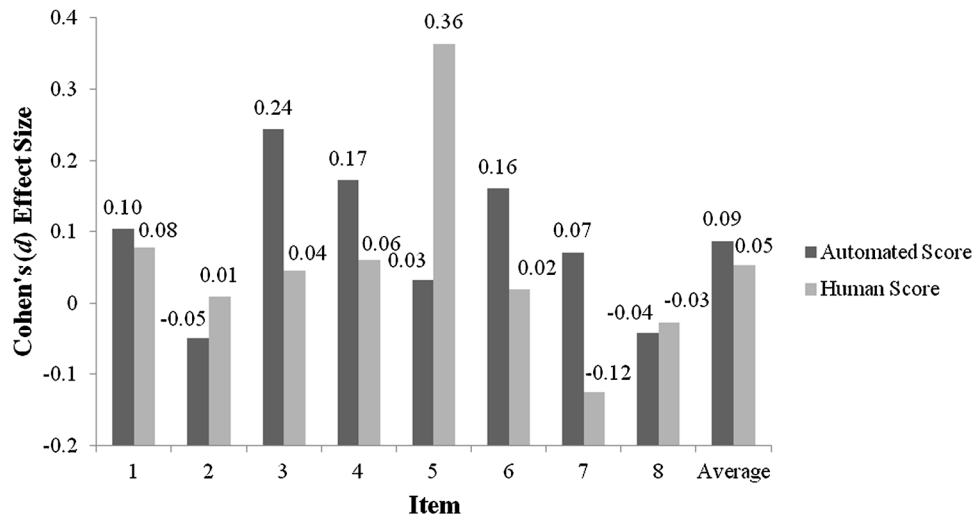


Figure 3. Effect size differences by language across scoring methods and items. The direction of the effect size results were based on subtracting the mean of non-native English speakers from those of native English speaker.

Analysis of Sources of Error

Across items, item 5 ($\kappa = 0.62$, $r = 0.66$) was shown to have relatively low agreement between human and automated ratings. Table 5 provides the scoring matrix for this item, which revealed a high discrepancy for scores of 2 or 3. More specifically, in nearly 13% of the ratings for item 5, the automated method gave a score of 3, while the human score was 2. Additionally, 13% of ratings were reversed with human raters giving a score of 3, while a score of 2 was given by the automated algorithm.

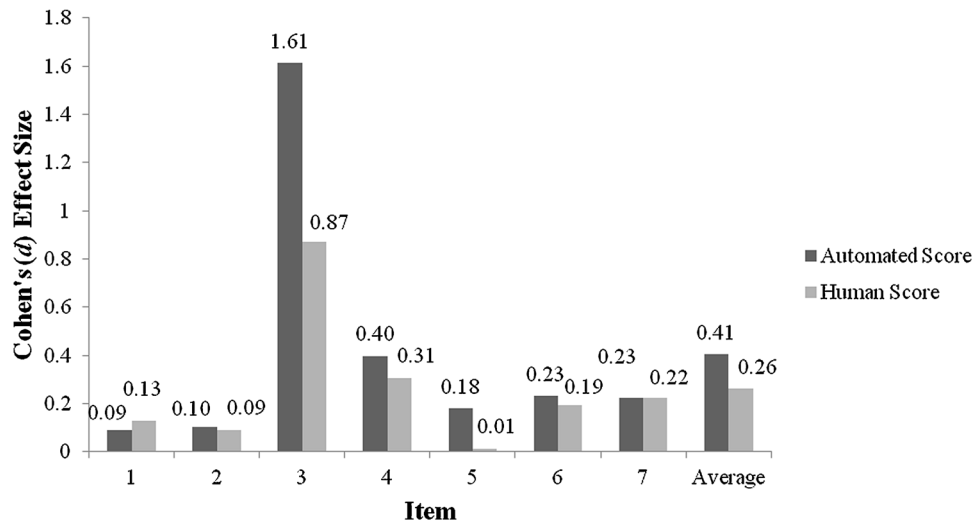


Figure 4. Effect size differences by computer use across scoring methods and items. This graph only shows results from seven items as the computer use information was missing from the last item. Furthermore, the direction of the effect size results were based on subtracting the mean of participants that did not have computer at home from those that did.

Table 5
Scoring matrix for item 5

Rating Human (Rows)	Automated Scores (Columns)					
	1	2	3	4	5	<i>n</i>
1	22	8	0	0	0	30
2	6	131	50	1	0	188
3	2	48	89	3	0	142
4	0	1	8	6	0	15
5	0	0	0	1	1	2
<i>n</i>	30	188	147	11	1	377

Note: Human and automated scores are the rows and columns respectively. Additionally, the scoring matrix is based on a proportion of the total ratings.

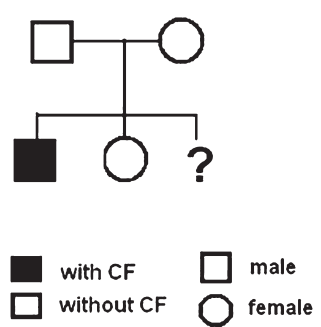
A plausible reason for the relatively low agreement between human and automated scoring may have been largely due to misspellings and linguistic diversity that was not modeled in c-rater-ML. Examples of score discrepancies for item five (see Fig. 5) are provided for illustrative purposes.

A sample response that illustrates misspellings that lead to differential scoring follows: *“its not on the dominant allelebut its on the resesive.the baby can still be born with it.”*

This response received a score of 3 by a human rater and a score of 2.3 by the automated algorithm which rounds to a score of 2. Within this example, there are two major spelling errors related to the correct response: one is that a space is missing between “allele” and “but,” and between “resesive.” and “the,” and the other is that “resesive” is a misspelling of “recessive.”

The character sequence features in c-rater-ML may help it detect the word “allele” from the “allelebut” response, but the word “recessive” may have gone undetected from the “resesive.the” response. This probably is not because a separation is missing between the end and beginning of the two sentences, as the period was most likely detected as a word-boundary and was not a major cause of concern. It is more likely that “recessive” was misspelled in such a way that the c-rater-ML model did not recognize it. In contrast, a human rater may have been able to infer that “resesive” was a misspelling of “recessive” and therefore gave the test taker credit for including this important term in the response.

An additional source of error may have come from linguistic diversity, which is illustrated in the following response:



Item 5. Cystic Fibrosis is a disease of the body’s sweat and mucous glands. It causes malnutrition, respiratory infections, and breathing difficulties. The picture (left) shows who has Cystic Fibrosis in a family with two children. The couple is expecting a third child. Will their baby be born with cystic fibrosis?

☐ Yes

☐ No

Figure 5. Examples of score discrepancies for item 5.

Table 6
Scoring matrix for item 7

Rating Human (Rows)	Automated Scores (Columns)					
	1	2	3	4	5	<i>n</i>
1	9	13	0	0	0	22
2	6	310	70	4	0	390
3	1	21	131	23	1	177
4	0	1	35	56	11	103
5	0	0	1	12	7	20
<i>n</i>	16	345	237	95	19	712

Note: Human and automated scores are the rows and columns respectively. Additionally, the scoring matrix is based on a proportion of the total ratings.

“Even though the parents do not have the disease, their DNA may still carry the disease, and that may actually pass on to their baby.”

This response was given a score of 4 by a human rater and a score of 2.98 by c-rater-ML. One potential explanation for the discrepancy between scores is the use of the word “DNA.” In particular, this word is being used as a synonym for genes in general and, as a result, may have not been well represented in the training responses for the automated algorithm, which did not consider “DNA” as a variant of “gene.” The human rater was most likely able to recognize this linguistic diversity, which led him/her to assign a higher score.

Besides misspelling and linguistic diversity issues, the discrepancies in scoring observed by the kappa and Pearson correlation values for this item may have been largely due to a restriction of range in scores. For instance, between automated and human raters as much as 87% of scores fell within the 2–3 range of the five-point scale for item 5 with relatively few responses for other score levels. For example, there were only four responses that had a score of 5 by human raters on this item (Table 5). As noted by Stemler (2004), restricted variability in the observed data will deflate consistency estimates such as kappa, which may lead to inaccurate appraisals of inter-rater agreement. As a result, items with similar misclassification rates but more evenly distributed responses across the score-range may have much higher overall agreement. For example, on item 7, nearly 13% of ratings were misclassified between categories 2 and 3 (Table 6). However, because of this item’s reasonable variability in observed scores (i.e., greater frequency of scores in the 4- and 5-point category), it yielded kappa and *r* values of 0.78 and 0.82, respectively. Such a result suggests that the low agreement values for item 5 may have largely been due to the restricted range of scores. It is important to note that interpreting the inner workings of these relatively complex NLP learning models is difficult, and so the explanations provided are plausible speculations.

Discussion

c-rater-ML’s Potential in Scoring Complex Science Items

These results illustrate the value of automated scoring of constructed response items and show the potential of c-rater-ML for complex science items. Prior studies have demonstrated the potential for automated scoring of science items for both large-scale and classroom assessments (Kuo & Wu, 2013; Linn et al., 2014; Liu et al., 2014; Nehm & Haertig, 2012) Consistent with previous research, overall c-rater-ML showed satisfactory results in scoring science items with five-point scoring rubrics as indicated by good to very good kappa values and high correlations

between human and automated scores. Subgroup analyses also revealed that in general there were no substantial differences between human and automated scores across gender, language, and computer use groups.

Compared to previously tested concept-based scoring engines such as c-rater (Liu et al., 2014), results from this study suggest that c-rater-ML offers an effective and efficient alternative. For c-rater scoring, the human rubric was transformed to an analytic rubric and the human/machine comparison was based on the analytic rubric, not on the human rubric. In contrast, c-rater-ML does not involve an analytic rubric and provides a direct comparison to the original human rubric. Thus, the c-rater-ML approach produces scores comparable to the human scores and omits the additional, cumbersome burden of the analytic rubric for human raters.

Sources of Error

The findings from this study may be affected by three possible sources of error in evaluating the adequacy of c-rater-ML: (1) sample size restrictions; (2) lack of human-human agreement statistics; and (3) possible unaccounted lexical and syntactic response variations.

In general, the overall sample size had a significant and large correlation with the accuracy of the automated scoring procedure ($r = 0.88$ between sample size and kappa and $r = 0.86$ between sample size and Pearson correlation), which was due largely to the increased variability in responses across the score distribution. Specifically, items with very few responses at specific score levels tend to have lower scoring consistency between human and automated scoring engines because the small set of responses does not allow the scoring engine to estimate appropriate values for the parameters used to map linguistic features to scores. Similar discussion was noted in Ha et al. (2011) when the authors included that low frequencies of certain responses contributed to less satisfactory scoring by automated engines. Machine scoring is typically challenging when the model is built on too few cases. Consistent with the nature of all automated scoring, small sample sizes and skewed distributions of score categories introduce challenges for c-rater-ML. This was likely the reason that items with more than 1,000 respondents had both quadratic-weighted kappas and Pearson correlations above 0.80, while all other items with less than 1,000 respondents had inter-rater agreement statistics below 0.80. Some of the items have a relatively small number of responses for modeling building (e.g., item 3). Increasing the sample size for some items could allow for improved evaluation of c-rater-ML.

Another source of error may relate to the accuracy of human scoring. The reliability of scores generated by human raters directly affects the quality of automated scores as the latter is evaluated against the former. In this study, two human raters rated a random subset of the responses ($n = 50$) for each item. The inter-rater reliability was 0.90 or above before the raters proceeded with individual scoring of the rest of the responses. Even though the overall human-human inter-rater agreement was lacking, the high agreement on the randomly selected responses reasonably suggests that the raters were sufficiently consistent in scoring.

Lexical and syntactic variations of responses (i.e., use of different words or grammatical structures to express similar ideas) that are not present in the model training sample and are therefore not captured by the scoring model are another likely source of error. The linguistic diversity will naturally affect scoring performance. For example, synonyms may not receive the same score because one word was well-represented in the training data and the other was not. Misspelling is another typical type of variation that may not be adequately modeled by scoring engines (Ha et al., 2011). Improvements in NLP are needed to deal with synonyms, misspellings, and odd grammatical structures. In the meantime, increasing the sample size for training models is likely to improve low human-machine agreement.

Implications for Science Education

Advances in automated scoring have significant implications for science education. Currently multiple-choice assessments are largely used in science classrooms and a partial reason is that it is challenging for teachers and instructors to spend the time scoring open-ended questions. Although multiple-choice items have their own merits, open-ended questions are needed to capture students' in-depth reasoning and argumentation (Haudek et al., 2012; Nehm et al., 2011; Nehm & Schonfeld, 2008). Automated scoring tools with proven accuracy can help increase the use of open-ended assessments in both K-12 and college-level science classrooms.

Another significant implication of machine scoring is that it offers the possibility of instant feedback to students. Feedback that points students to specific instructional steps can be designed to be associated with each automated score. For example, Linn et al. (2014) used c-rater scores to assign guidance and reported that immediate automated guidance was as effective as delayed teacher guidance in prompting students to review, revise, and improve their responses. At an aggregate level, automated scores and feedback can inform teachers of the common barriers to understanding and facilitate teachers' timely instructional adjustments. Additionally, given the objectivity of machine scores, they can facilitate comparison of students' science performance across instructors and teachers (Weston et al., 2013).

Future Directions of Research

Next steps for research on automated scoring of constructed response items include clarifying the factors that contribute to high agreement. For example, future research could further clarify the relationship between sample size and accuracy of automated scoring. Simulation studies can be conducted to identify thresholds of sample sizes associated with various levels of scoring accuracy. The simulation should also factor in the distribution of responses across score levels to determine how the final results are influenced by score distribution. Results from such investigations can provide helpful guidelines in terms of the selection of items for automated scoring.

Testing the practical validity of automated scores for use in classroom science teaching is another direction for future research. Scores can be used to diagnose student misconceptions and provide guidance, for example. Further research is required to advance understanding of when, how, and under what circumstances automated feedback works best to inform science instruction and improve science learning. In the case of c-rater-ML, the advantage of possessing both a faster model building process and more accurate scoring when compared to concept-based scoring approaches has the potential to assign guidance for formative items embedded in science inquiry units.

Another direction for research on classroom uses of automated scores could be to differentiate among responses within a score category. For example, in the knowledge integration scoring rubric (Table 3), a response gets a score of 2 if it represents a scientifically invalid idea. In most cases there are several categories of invalid ideas that all receive the same score. Therefore, in the future, c-rater-ML scoring can be designed to capture the nuances across different responses that receive the same score. Such differentiation in automated scores would allow the guidance to be designed to target the specific kind of invalid idea associated with each type of score, which has the potential to assist students in improving their science understanding.

A third direction for classroom use is to use automated scores to provide snapshots of class progress to the teacher. Teachers could then use this specific information to plan lessons aligned with the progress of their students.

Limitations

A limitation of this study, as mentioned earlier, is that the overall human inter-rater agreement was not available. However, the high human agreement (e.g., 0.90 or above) on a random subset of the responses suggests that it is reasonable to assume overall satisfactory consistency in human scoring. Another potential limitation is that no cognitive interviews were conducted with students to understand their underlying thinking with their written responses to these items. Such interviews may provide useful information for the discrepancy analysis.

Conclusion

This study provides one of the first sets of empirical evidence for the use of c-rater-ML, an NLP tool designed for scoring short-response science items. This tool is more efficient and more aligned with human scoring than some of the previously tested concept-based scoring tools, and yields satisfactory agreement with human scores. The accuracy of the c-rater-ML demonstrated in this study is also consistent with that of the high performing automated scoring engines used in prior studies (see Table 1). Therefore, c-rater-ML has shown great potential to score complex science assessments that measure understanding on inquiry problems. Our next steps will be to investigate how this automated scoring tool can be used to facilitate instant feedback to both students and teachers for improved learning and teaching of science topics.

Author Note

This material is based upon work supported by the National Science Foundation under Grant No. 1119670. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Note

¹ c-rater-ML is not an open-source scoring engine. However, external researchers could have access to c-rater-ML scores through research collaborations with Educational Testing Service.

References

- Attali, Y., & Powers, D. (2008). Effect of immediate feedback and revision on psychometric properties of open-ended GRE R subject test items (GRE Board Research Rep. No. 04-05; ETS RR-08-21). Princeton, NJ: Educational Testing Service.
- Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance?. *Journal of Science Education and Technology*, 23, 160–182.
- Bennett, R. E., & Sebrechts, M. M. (1996). The accuracy of expert-system diagnoses of mathematical problem solutions. *Applied Measurement in Education*, 9, 133–150.
- Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, 25, 27–40.
- Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers (PDF). In *Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing*. College Park, MD.
- Burstein, J., & Marcu, D. (2002). Automated evaluation of discourse structure in student essays. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 200–219). Mahwah, NJ: Lawrence Erlbaum.
- Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27, 355–377.

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e-rater's performance on TOEFL essays (TOEFL Research Report No. RR-73; ETS RR-04-04). Princeton, NJ: Educational Testing Service.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: Erlbaum.

Dzikovska, M., Moore, J. D., Steinhäuser, N., Campbell, G., Farrow, E., & Callway, C. B. (2010). Beetle II: A system for tutoring and computational linguistics experimentation. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 13–18). Uppsala, Sweden: Association for Computational Linguistics.

Dzikovska, M. O., Nielsen, R. D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., . . . Dang, H. T. (2013). SemEval-2013 Task 7: The joint student response analysis and 8th recognizing textual entailment challenge. *Proceedings of the Seventh International Workshop on Semantic Evaluation: Vol. 2. Second Joint Conference on Lexical and Computational Semantics* (pp. 263–274). Atlanta, GA: Association for Computational Linguistics.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33, 613–619.

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

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, 1, 3–31.

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. *CBE-Life Sciences Education*, 10, 379–393.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE-Life Sciences Education*, 11(3), 283–293.

Heilman, M., & Madnani, N. (2013). ETS: Domain adaptation and stacking for short answer scoring. *Proceedings of the 2nd joint conference on lexical and computational semantics: Vol. 2. Seventh International Workshop on Semantic Evaluation* (pp. 275–279). Atlanta, GA: Association for Computational Linguistics.

Higgins, D., Zechner, K., Xi, X., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, 25, 282–306.

Kuo, C. Y., & Wu, H. K. (2013). Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science. *Computer & Education*, 68, 388–403.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37, 389–405.

Lee, H. S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education*, 24, 115–136.

Linn, M. C., & Eylon, B.-S. (2006). Science education: Integrating views of learning and instruction. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 511–544). Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, M. C., Gerard, L., Ryoo, K., McElhaney, K., Liu, O. L., & Rafferty, A. N. (2014). Computer-guided inquiry to improve science learning. *Science*, 344(6180), 155–156.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, 33, 19–28.

Liu, O. L., Lee, H. S., Hoftstetter, C., & Linn, M. C. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment*, 13, 33–55.

Liu, O. L., Lee, H. S., & Linn, M. C. (2011). Measuring knowledge integration: Validation of four-year assessments. *Journal of Research in Science Teaching*, 48, 1079–1107.

Liu, O. L., Ryoo, K., Sato, E., Svihla, V., & Linn, M. C. (2015). Designing assessment to measure cumulative learning of energy topics. *International Journal of Science Education*. DOI: 10.1080/09500693.2015.1016470

Mayfield, E., & Rosé, C. (2010). An interactive tool for supporting error analysis for text mining. Proceedings of the demonstration session at the International Conference of the North American Association for Computational Linguistics (pp. 25–28). Los Angeles, CA: Association for Computational Linguistics.

Mitchell, T., Russell, T., Broomhead, P., & Aldridge, N. (2002). Towards robust computerized marking of free-text responses. Proceedings of the International Computer Assisted Assessment Conference (pp. 233–249). Loughborough, UK: Loughborough University.

Moharrer, K., Ha, M., & Nehm, R. (2014). EvoGrader: An online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, 7(1), 1–15. Retrieved from <http://www.evolution-outreach.com/content/7/1/15>

Mohler, M., Bunesco, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Vol. 1. Human Language Technologies (pp. 752–762). Portland, OR: Association for Computational Linguistics.

National Council of Teachers of English. (2008). Statement on class size and teacher workload: Secondary. Retrieved from <http://www.ncte.org/positions/statements/classsizesecondary>.

Nehm, R. H., Ha, M., & Mayfield, E. (2011). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, 21(1), 183–196. doi: 10.1007/s10956-011-9300-9

Nehm, R. H., & Schonfeld, I. S. (2008). Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching*, 45, 1131–1160.

Nehm, R. H., & Haertig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: Testing the efficacy of text analytic software. *Journal of science education and technology*, 21(1), 56–73.

Nielsen, R. D., Ward, W., & Martin, J. H. (2008). Classification errors in a domain-independent assessment system. Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (pp. 10–18). Stroudsburg, PA: Association for Computational Linguistics.

Sandene, B., Horkay, N., Bennett, R., Braswell, J., & Oranje, A. (2005). Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project, research and development series (NCES 2005-457). Washington, DC: U.S. Government Printing Office.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Journal of Statistics and Computing*, 14(3), 199–222.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4), 66–78.

Sukkarieh, J. Z., & Bolge, E. (2010). Building a textual entailment suite for evaluating content scoring technologies. Proceedings of the Seventh Conference on International Language Resources and Evaluation (pp. 3149–3156). Paris, France: European Language Resources Association.

Sukkarieh, J. Z., & Blackmore, J. (2009). c-Rater: Automatic content scoring for short constructed responses. In H. C. Lane & H. W. Guesgen (Eds.), Proceedings of the twenty-second international Florida artificial intelligence research society conference (pp. 290–295). Menlo Park, CA: Association for the Advancement of Artificial Intelligence Press.

Weston, M., Parker, J.M., & Urban-Lurain, M. (2013). Comparing formative feedback reports: Human and automated text analysis of constructed response questions in biology. Paper presented at the Annual Conference of the National Association on Research in Science Teaching, Rio Grande, Puerto Rico.

Williamson, D., Xi, X., & Breyer, J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13.