Title
Automated scoring of constructed-response science items: Prospects and obstacles

Journal
Educational Measurement: Issues and Practice, 33(2)

ISSN
0731-1745

Authors
Liu, OL
Brew, C
Blackmore, J
et al.

Publication Date
2014

DOI
10.1111/emip.12028

Peer reviewed

Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Summer 2014, Vol. 33, No. 2, pp. 19–28

Automated Scoring of Constructed-Response Science Items: Prospects and Obstacles

Ou Lydia Liu, *Educational Testing Service,* Chris Brew, *Nuance,* John Blackmore, *Educational Testing Service,* Libby Gerard, Jacquie Madhok, and Marcia C. Linn, *University of California, Berkeley*

Content-based automated scoring has been applied in a variety of science domains. However, many prior applications involved simplified scoring rubrics without considering rubrics representing multiple levels of understanding. This study tested a concept-based scoring tool for content-based scoring, c-rater™, for four science items with rubrics aiming to differentiate among multiple levels of understanding. The items showed moderate to good agreement with human scores. The findings suggest that automated scoring has the potential to score constructed-response items with complex scoring rubrics, but in its current design cannot replace human raters. This article discusses sources of disagreement and factors that could potentially improve the accuracy of concept-based automated scoring.

Keywords: automated scoring, constructed-response items, c-rater™, science assessment

Constructed-response items are an important tool to elicit students' in-depth understanding of science content and measure students' ability to communicate scientific ideas. Both science education researchers (e.g., Lane, 2004; Shepard, 2000) and the Next Generation Science Standards (NGSS; National Research Council, 2012) stress that scientific explanation is a critical component of science education. As the NGSS points out, multiple-choice items can provide snapshots of some types of science knowledge, but open-ended assessment items are essential for measuring other competencies such as the ability to formulate a problem, conduct investigations, and communicating findings. Yet, the costs associated with scoring student constructed responses often stand in the way of their use. Wainer and Thissen (1993) estimated that the scoring of 10 constructed-response items costs about $30 while the scoring of multiple-choice items to achieve the same level of reliability costs only about 1¢.

To increase the use of constructed-response items, automated scoring has been explored over the past two decades in a variety of domains. If accurate, automated scoring can reduce the interval between test administration and score assignment, reduce the number of human scorers, save on costs for training human scorers, and potentially improve the scoring consistency (Burstein & Marcu, 2002; Burstein, Marcu, & Knight, 2003; Williamson, Xi, & Breyer, 2012).

Despitethe wide applications of automated scoring, many prior applications focus on simplified scoring rubrics (e.g., containing three levels: correct, partially correct, or wrong), without considering items involving multiple levels of understanding. Using a concept-based scoring tool, c-rater™, this article explores automated scoring of constructed-response science items with rubrics representing multiple levels of understanding. Our goal was to contribute to the understanding of the prospects and obstacles of using automated scoring with complex rubrics. We hope to identify the conditions when automated scoring can produce reliable scores and make constructed-response items more accessible to science researchers and teachers (Quellmalz & Pelligrino, 2009).

Research on Content-Based Automated Scoring in Educational Contexts

Automated scoring has been widely applied in educational research to improve scoring efficiency and shorten the time between test administration and when teachers, test takers, and score users receive test results. Research on automated scoring has covered domains such as writing quality (Burstein & Marcu, 2002; Foltz, Laham, & Landauer, 1999), mathematics (Bennett & Sebrechts, 1996; Sandene, Horkay, Bennett, Braswell, & Oranje, 2005), written content (Attali & Powers, 2008; Dzikovska, Nielsen, & Brew, 2012; Graesser, 2011; Leacock & Chodorow, 2003; Mitchell, Russell, Broomhead, & Aldridge, 2002; Nielsen, Ward, & Martin, 2008; Sukkarieh & Bolge, 2010), speech (Bernstein, Van Moere, & Cheng, 2010; Higgins, Zechner, Xi, & Williamson, 2011), and

Ou Lydia Liu, Educational Testing Service, 660 Rosedale Road, MS07-R, Princeton, NJ 08541; lliu@ets.org. Chris Brew, Nuance, 1198 East Arques Avenue, Sunnyvale, CA 94085; cbrew@acm.org. John Blackmore, Educational Testing Service, 660 Rosedale Road, MS75-D, Princeton, NJ 08541; jblackmore@ets.org. Libby Gerard, University of California Berkeley, Tolman Hall, Berkeley, CA 94720; libbygerard@berkeley.edu. Jacquie Madhok, University of California Berkeley, Tolman Hall, Berkeley, CA 94720; jjmadhok@gmail.com. Marcia C. Linn, Graduate School of Education, University of California Berkeley, 4523 Tolman Hall, Berkeley, CA 94720; mclinn@berkeley.edu.

other education related topics (Sargeant, Wood, & Anderson, 2004).

In the following section, we focus our review of automated scoring of content-based constructed responses. Content-based scoring refers to the type of scoring that evaluates the content of the responses, as opposed to the writing quality (e.g., essay). Concept-based scoring is one type of content-based scoring. There are other types of content-based scoring that do not involve the scoring of individual concepts. Most previous studies used the following criteria to evaluate the accuracy of c-rater scoring: quadratic-weighted kappa (Fleiss & Cohen, 1973), Pearson correlation, degradation of human/machine agreement from human/human agreement, and standardized mean score differences between automated and human scores (Williamson et al., 2012). We first introduce these criteria for automated scores so the readers understand the context of the literature review.

Evaluation Criteria

Quadratic-Weighted Kappa

The kappa coefficient indicates the proportion of agreement between two raters beyond what is expected by chance and is scaled to range from −1 to 1, with −1 indicating poorer than chance agreement, 0 indicating pure chance agreement, and 1 indicating perfect agreement. Kappa was subsequently developed into the quadratic-weighted kappa coefficient that can be applied to multiple raters (Fleiss & Cohen, 1973).

The choice of threshold values for the kappa coefficient is rather arbitrary. Williamson et al. (2012) argued that a satisfactory kappa coefficient for automated scores should be at least .70 to be used in high-stakes testing situations. Landis and Koch (1977) proposed the following as standards for the strength of agreement for the kappa coefficient: poor (≤.00), slight (.00–.20), fair (.21–.40), moderate (.41–.60), good (.61–.80), and very good (.81–1). Given that all the items in this study are used for low-stakes purposes and also that the Landis and Koch standards offer more detailed distinction among kappa categories, we adopted these standards for this study.

Pearson Product-Moment Correlation

Pearson correlation has been used as a common criterion to evaluate the consistency between human and machine scores. The interpretation of the Pearson correlation is also arbitrary, as is the case for kappa. We followed the Cohen (1968) rules for classification: none (0–.09), small (.10–.30), moderate (.31–.50), and large (.51–1.00).

Degradation from the Human/Human Score Agreement

This criterion examines the difference between human/machine score agreement and human/human score agreement (e.g., if human/human agreement is .80 in kappa and human/machine agreement is .60 in kappa, then the degradation is .20 in kappa). Williamson et al. (2012) proposed that the degradation of automated scoring agreement from human agreement should not be greater than .10 in either kappa or Pearson correlation.

Standardized Mean Score Difference

The standardized mean score difference between the automated and human scores is another evaluation criterion. Essentially, it is the mean difference between human and machine scores divided by the pooled standard deviations. The standardized mean score difference should not be greater than .15 (Williamson et al., 2012).

Prior Research on Content-Based Scoring Tools

Over the past 20 years, content-based scoring has been explored in various domains (Table 1). c-Rater has been tested in a number of subject domains with an array of test taker populations (Sukkarieh & Bolge, 2010; Sukkarieh & Pulman, 2005). It also has been tested on large-scale assessments, such as the NAEP ICT Science (e.g., Leacock & Chodorow, 2003; Sukkarieh, Pulman, & Raikes, 2003). For example, Leacock and Chodorow (2003) used c-rater to evaluate constructed-response math reasoning and reading comprehension items, with average quadratic-weighted kappa being .73 for both domains. Furthermore, c-rater was used to score items adapted from the Graduate Record Examination (GRE®) subject tests in biology and psychology (Attali, Powers, Freedman, Harrison, & Obetz, 2008). Based on the three-level rubrics, the authors reported that the average quadratic-weighted kappa between human and c-rater scores was .62 for biology items and .83 for psychology items. The authors also looked at the accuracy of feedback assignment by machine and humans, with the mean quadratic weighted kappa being .57 for biology items and .81 for psychology items.

In addition to c-rater, there are other automated scoring tools designed to score content-based constructed-response items (Dzikovska et al., 2012; Graesser, 2011; Graesser, Rus, D'Mello, & Jackson, 2008; Pulman & Sukkarieh, 2005; VanLehn et al., 2007; Wiemer-Hastings, Arnott, & Allbritton, 2005). For example, the Oxford-UCLES system (Pulman & Sukkarieh, 2005) draws on several machine learning techniques such as inductive logic programming, decision tree learning, and Bayesian learning in categorizing students responses into prespecified classes. The AutoTutor system was designed as an intelligent tutoring environment, originally in the areas of basic computer science and Newtonian physics (Graesser, 2011). Many of its spin-off versions have been developed to accommodate automated scoring of tasks in many other subject areas and skills such as critical thinking (Graesser, 2011; VanLehn et al., 2007). Dzikovska et al. (2010) tested automated scoring of college-level physics items and reported the machine/human agreement in kappa being .69. Nielsen et al. (2008) tested automated scoring of science items among for third to sixth graders and obtained a machine/human agreement in kappa of .73. Similarly, Bennett and Sebrechts (1996) evaluated the accuracy of an automated scoring system for scoring algebra word problems among GRE test takers. The automated scores agreed highly with human scores (i.e., 91%) in evaluating the correctness of the responses. However, the agreement was less satisfactory (i.e., 71%) when the evaluation was decomposed to the individual errors identified in the response. Automated scoring also showed potential in scoring problem-solving ability in earth science contexts, with Pearson correlations ranging from .67 to .82 (Wang, Chang, & Li, 2005), indicating good agreement. Machine scoring of students' summary of texts

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	Evaluation Criterion	
							M	*SD*
Attali et al. (2008)	c-Rater	College	331	Biology	3	1–3 sentences	.62[a]	.13
			640	Psychology	3	1–3 sentences	.83[a]	.10
			331	$Biology_{feedback}$	–	–	.57[a]	.14
			640	$Psychology_{feedback}$	–	–	.81[a]	.13
Bennett & Sebrechts (1996)	GIDE Algebra	Graduate	60	Algebra	2	–	.91[b]	–
Dzikovska et al. (2010)	BEETLE II	College	73	Physics	5	1–2 sentences	.69[a]	–
Kintsch et al. (2000)	Summary Street	Grade 6	39–56	Energy, Ancient Civilization, Circulatory System	4	75–300 words	.61[c]	.28
Leacock & Chodorow (2003)	c-Rater	Grade 4	245–250	Math	2–5	15 words	.73[a]	.02
	c-Rater	Grade 8	245–250	Math	2–5	15 words	.72[a]	.09
	c-Rater	Grade 11	16,625	Reading	3	43 words	.73[a]	.05
Nielsen et al. (2008)	SCIENTSBANK	Grades 3–6	16,000	Science	4	–	.73[a]	–
Wang et al. (2005)	UPSAM	Grade 10	20	Earth Science	–	–	.73[b]	.56

Note. [a]Kappa; [b]% of absolute agreement; [c]Pearson correlation.

on ancient civilizations correlated about .64 with a human rater, comparable to the correlation of .69 between two human raters (Kintsch, Steinhart, Stahl, Matthews, & Lamb, 2000).

In summary, content-based automated scoring demonstrates promise for certain constructed-response items in that the automated scores showed good to high agreement with human scores (e.g., Attali et al., 2008; Bennett & Sebrechts, 1996; Wang et al., 2005). However, many previous studies used simplified, up-to-three-level rubrics in scoring (e.g., Attali et al., 2008; Leacock & Chodorow, 2003). Among the applications that used more than three scoring levels, the levels either represent a mechanical composition of the correct answer (e.g., partial scores assigned when one part of the correct answer is missing; Dzikovska et al., 2010), or are not clearly defined (e.g., in Kintsch et al. [2000], the four levels of content scoring are described as good, OK, needs improving, or missing). It is important to have a clearer understanding of how content-based scoring performs when rubrics representing multiple levels of understanding are involved.

Research Questions

Our goal was to investigate the prospects and obstacles of c-rater scoring when items require rubrics that capture multiple levels of understanding. We addressed the following questions:

(1) Can holistic scoring rubrics be transformed to concept-based, analytic scoring rubrics for automated scoring? What is the correlation between the automated, analytic scores and the holistic human scores?
(2) Can concept-based automated scoring accurately score science explanation items with rubrics representing multiple levels of understanding?
(3) What are the main sources of disagreement in developing automated scores for explanation items?

Methods

In this study, we adopted the following steps in implementing the automated scoring process: item selection, transformation of holistic rubrics to concept-based analytic rubrics, human scoring using the c-rater analytic rubrics, c-rater scoring, statistical analyses, and model refining. This process took place over an 8-month period and involved an interdisciplinary team consisting of two automated scoring scientists, three content experts, one automated scoring engineer, and one measurement expert.

c-Rater Characteristics

c-Rater evaluates performance based on a set of clear, distinct concepts (Sukkarieh & Blackmore, 2009). The accuracy of c-rater scores depends on the linguistic complexity and the cognitive skills exhibited in the responses. Implementing c-rater involves four major steps: (1) *model building,* by which researchers identify one or more model responses that contain key concepts of the item; (2) *natural language processing* (NLP), by which student and model responses are analyzed for linguistic features using NLP methods or knowledge representation methods; (3) *main points identification*, by which linguistic features are used to determine the absence or presence of key concepts in the student responses; and (4) *scoring*, by which a score is assigned to a response based on prespecified scoring rules. c-Rater applies a sequence of NLP steps including correcting students' spelling, determining the grammatical structure of each sentence, resolving pronoun reference, and analyzing paraphrases in student responses. The depth of the linguistic analyses is intended to prevent

Table 2. The Four Items Tested with Automated Scoring

Item 1 (Sun): Explain how the sun helps animals survive.

Item 2 (Spoon): A metal spoon, a wooden spoon, and a plastic spoon are placed in hot water. After 15 seconds which spoon will feel hottest?

A. The metal spoon
B. The wooden spoon
C. The plastic spoon
D. The three spoons will feel the same.

Explain your choice.

Item 3 (Coal): 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. A warmer climate
B. A cooler climate
C. Lower relative humidity
D. More ozone in the atmosphere

Explain your choice.

Item 4 (Heat): In general, are heat energy and temperature the same or different?

Circle one Same Different

What is the main reason for their similarity or difference?

c-rater from being misled by responses that use the right words in the wrong contexts.

Item Selection

The items used in this study were designed to measure students' ability to explain science phenomena using coherent evidence (Liu, Lee, Hofstetter, & Linn, 2008; Liu, Lee, & Linn, 2010; Lee & Liu, 2010). These items emerged from research showing that students develop many varied ideas about science topics as a result of deliberate investigations of the natural world (e.g., the sun warms the air so plants can grow), culturally mediated conversations with family members (e.g., plants need food), and interpretations of everyday events (e.g., plants eat dirt). The items require students to develop an argument to support their explanation of a scientific phenomenon.

Four constructed-response items were selected for scoring with c-rater based on the availability of scored responses and centrality to instruction content. All four items were used in middle school classrooms where students had previously participated in another National Science Foundation (NSF)–supported project. Items came from the start-of-year or end-of-year assessments, unit pre/post tests, or embedded assessment within a unit. They included one stand-alone constructed-response item (*sun*; see Table 2) and three constructed-response items that are a follow-up to a preceding multiple-choice question (*spoon, coal, and heat*). The items were designed to measure student understanding of energy source, energy transfer, and energy transformation. The items address different science concepts and courses: *sun* targets photosynthesis in life science for seventh grade; *spoon* targets thermodynamics in physical science for sixth grade; *coal* targets global climate change in earth science for sixth grade; and *heat* targets thermodynamics in physical science for sixth grade. The four items were scored by human raters and demonstrated satisfactory psychometric properties such as item fit, item discrimination, and point-biserial correlation (Liu, Lee, & Linn, 2011). The number of available student responses ranged from 475 to 550 for each of the four items.

The Five-Level Holistic Scoring Rubric

All four items were scored by human raters using holistic 5-point scoring rubrics previously developed by Linn and Eylon (2011). The scoring levels were *off task*, *no link*, *partial link*, *full link*, and *complex link*, representing progression from invalid to complex understanding (Lee, Varma, Linn, & Liu, 2010). Off-task answers are typically blank answers or answers such as "I don't know." No-link answers are responses with scientifically invalid ideas. Partial link refers to responses showing some evidence of valid ideas but in need of further elaboration to demonstrate understanding. At the full-link level, responses include scientifically valid ideas and show evidence of students' ability to see connections between ideas. At the highest complex-link level, students are able to use multiple pieces of evidence in explaining science phenomena and articulate the connections between the evidence. Another common feature of these advanced responses is that students use proper scientific language in their explanation. The scoring rubrics were similar across all the items but featured customized definitions and exemplars for each individual item. (See Table 3 for a sample rubric with the five levels for the *spoon* item.) The following is an example of the complex-link level response for Item 1 (*sun*; Table 2), "Explain how the sun helps animals survive":

> The sun helps animals survive by giving light energy to plants. Light energy is needed in photosynthesis. Plants can take the process of photosynthesis to result in glucose. Glucose is sugar that helps the plant grow. So when an animal eats a plant the energy is transferred to the animal. The energy helps make the animal live.

The holistic rubric has shown to be effective in differentiating among students of various ability levels (Liu et al., 2011). Students scoring high on a particular item also tend to have higher average scores, considering all other items (Liu et al., 2008, 2011).

Typically for human scoring, two raters rated about 10% of the responses for each item to reach a .90 quadratic-weighted kappa agreement. Scores were reconciled for responses that showed any rating discrepancies.

Table 3. Holistic Scoring Rubric for the *Spoon* Item

Score	Level	Description	Example
1	Off-task	No answer or off-task	"I don't really know."
2	No link	Nonnormative or scientifically invalid links and ideas	"The metal spoon traps heat the best and will stay hot longer." "Because when a metal spoon get hot it stays hot for a little while."
3	Partial link	Normative ideas without scientifically valid connections between ideas	"The metal spoon because metal heats up very much in a small amount of time." (Metal is not compared with other material.)
4	Full link	One scientifically valid and elaborated link between normative and relevant energy ideas	"Metal transfers heat faster than plastic or wood."
5	Complex link	Two or more scientifically valid and elaborated links between normative and relevant energy ideas	"All of the spoons will be the same temperature, but the metal spoon will feel the hottest because it is a better conductor than plastic or wood."

Development of the c-Rater Analytic Rubric and Scoring Rules

Development of the c-rater analytic rubric required transforming the holistic rubric into a set of concepts that could be combined to form a score; the analytic rubric needs to capture the knowledge and understanding represented in the original rubric. The analytic rubric differs from the holistic rubric in that it breaks the holistic evaluation of scientific explanation into the important concepts or main ideas assessed by each item. The analytic rubric also needs to identify the alternative expressions of the concepts that students could possibly provide in their responses.

The first versions of the analytic rubrics were developed by the content experts. They were then reviewed by two NLP scientists, a c-rater engineer, and a measurement expert. The reviews mainly focused on the scientific ideas represented by the identified concepts, the inclusiveness of the alternative expressions, the clarity of the wording, and any notes that should be provided that are not included in the concepts. The final analytic rubric was then used by both c-rater and two human raters to score student responses.

A challenge in developing the analytic rubrics was to strike a balance between including all concepts necessary for alignment with the holistic rubrics and keeping the number of concepts manageable for human raters to use. For this reason, we took an iterative approach in developing the analytic rubrics. The first drafts of analytic rubrics had large numbers of concepts, ranging from 12 to 18 for each of the four test items. When human raters tried to score the items using these rubrics, the interrater agreement was very low (e.g., .33–.52 in quadratic weighted kappa), which suggests that, although these rubrics captured most of the important aspects of the holistic rubrics, they were too cumbersome for human raters to assign scores reliably. For each response, the human raters needed to check the response against all possible concepts and their alternative expressions to determine the presence or absence of a concept. To place the workload into context, to score an item with 400 responses using an analytic rubric with 18 concepts, the human raters need to rate the combination of responses and concepts 7,200 (i.e., 400*18) times for only one item! With the large number of concepts, it was infeasible for human raters to achieve satisfactory agreement, a critical step in automated scoring implementation, because as human scores are typically used as the reference to evaluate the accuracy of machine scores.

To alleviate this problem, we consolidated the concepts to maximize information extraction while reducing the complexity for human scoring. The final rubrics had 6–10 concepts for each of the four items. As an illustration, the rubric for the *spoon* item has six main concepts (see the appendix). The five-level holistic scoring rubric was then transformed into a four-level scoring rule, with the analytic score 4 corresponding to the holistic full-link responses or above (holistic scores 4 and 5 were combined because not many scores of 5 were observed in the responses), 3 to the partial-link level, 2 to the no-link level, and 1 to the off-task level.

Each concept has a number of paraphrases. In addition to specifying the paraphrases, rubric developers added notes about acceptable or unacceptable similar words. For example, under Concept 5 for the *spoon* item, the unacceptable substitutes for *attract* include *absorb*, *take in*, *extract*, *transfer*, and *conduct*. After the analytic rubrics were finalized, scoring rules were created for each item. The scoring rules specify the various combinations of the presence of concepts (see the appendix).

Human Rater Training and Scoring

We identified a group of postdoctoral researchers, research scientists, and advanced graduate students to be trained to score the items using the analytic rubric. All of the raters had a college-level science background.

To score each response, raters checked the student response against each of the c-rater concepts. The raters highlighted each concept and indicated whether this concept was present or absent in the response being evaluated. For training purposes, the two raters first used 25% of the responses to become familiar with the scoring procedure and to establish agreement in scoring. After the raters indicated that they felt comfortable with the scoring, they then scored the remaining 75% of the responses. Therefore, all the findings reported in the Results section refer to 75% of the responses. Raters reported that they typically spent one to three minutes scoring each response, depending on the number of concepts that they needed to check for each item and their familiarity with the scoring platform.

Table 4. Descriptive Statistics of the Human and c-Rater Scores

	N		H1		H2		AH		c-Rater	
	75%[a]	Blind Evaluation	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Sun	412	103	2.18	1.14	2.39	.90	2.25	.92	2.77	.88
Spoon	362	90	2.01	.89	1.90	.89	1.96	.86	1.84	.85
Coal	321	80	1.31	.56	1.32	.54	1.31	.55	1.18	.42
Heat	356	89	1.39	.68	1.25	.57	1.32	.59	1.25	.51

Note. [a]The column reflects data from 75% of the total responses (see Human Rater Training and Scoring section). From this subset, half of the responses were used for model development, 25% were used for cross validation, and 25% were used for blind evaluation.
H1 = human rater 1; H2 = human rater 2; AH = average of H1 and H2.

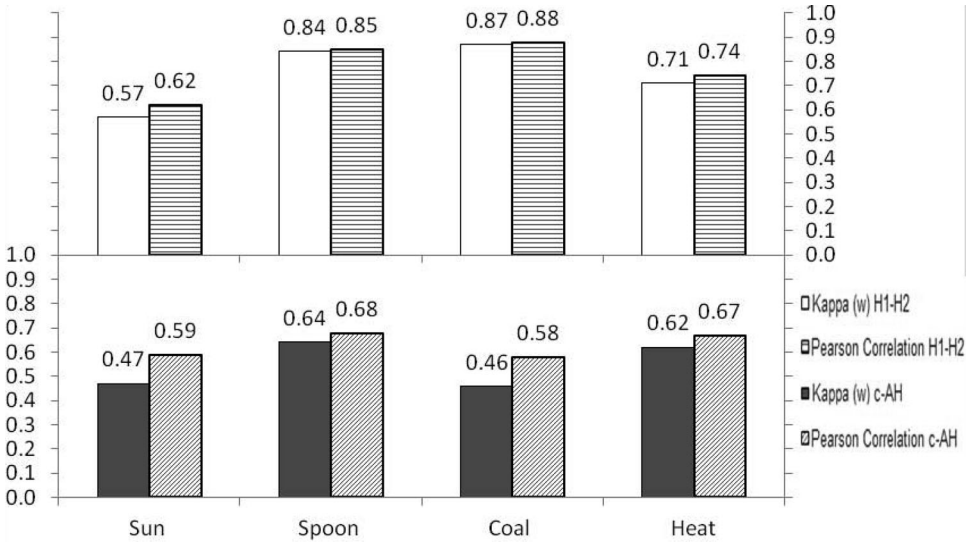

FIGURE 1. Quadratic-weighted kappa and Pearson correlation between scores from human raters 1 and 2, and between the average of human raters and c-rater scores. H1 = human rater 1; H2 = human rater 2; c = c-rater; AH = average between the two human raters.

c-Rater Model Building

After human raters completed their scoring, we moved on to c-rater model building. Before a c-rater scoring model is developed, all available annotated response data were randomly divided into three sets: the development, validation, and blind sets, using the ratio 2:1:1, respectively. The c-rater model was derived from an analytic rubric and the annotated development data set. In many cases, additional scoring notes were provided by the item developers to inform the model building process. For example, information indicating which words were significant to the prompt and any similar words or phrasings that may be accepted, was critical for this process. The validation set was used as a predictor of c-rater's performance on unseen, or blind, response data, and informed the model builder of any adjustments that may have been needed, judging from the increase or decrease of scoring agreement with human raters. Because the model was often updated based on its performance on the validation set, an additional blind set was required for evaluation purposes.

In this study, we used the evaluation criteria described in the prior research section, including quadratic-weighted kappa, Pearson correlation, degradation from human/human agreement, and standardized mean difference. In addition to the above criteria, we also looked at the correlation between human/c-rater scores and the number of words in the responses. We expected the correlations to be low, because both human and c-rater scores should be determined by the underlying concepts captured in the responses, not the length of the response.

Table 5. Degradation of c-Rater/Human Agreement from Human/Human Agreement and Standardized Mean Difference

	Degradation		Standardized Mean Difference
	Kappa	Pearson Correlation	(AH–c-Rater)
Sun	−0.10	−0.03	−0.57
Spoon	−0.20	−0.17	0.12
Coal	−0.31	−0.30	−0.25
Heat	−0.09	−0.07	0.13

Note. AH = average between H1 and H2.

Results

Comparison of Human and c-Rater Scores

When comparing human and c-rater scores, we found that for three of the four items, human raters tended to assign a higher score than c-rater (Table 4). Comparing human raters 1 and 2 with c-rater on the quadratic-weighted kappa and Pearson correlation criteria reveals moderate to good agreement (Figure 1). The items showed large Pearson correlations (Cohen, 1968) and moderate to large-kappa-values (Landis & Koch, 1977) between human and machine scores. However, the human/machine consistency was considerably better on some items (i.e., *spoon* and *heat*) than others (i.e., *sun and coal*). Items *spoon and coal* showed unacceptable degradation (i.e., >.10) from human score agreement (Table 5). Items *sun* and *heat* showed unacceptable

Table 6. Pearson Correlation Between c-Rater Scores and Human Scores Based on the Original Holistic Rubric

	N	Correlation
Sun	412	.67
Spoon	362	.72
Coal	321	.68
Heat	356	.70

standardized mean difference (i.e., >.15). There was clearly an interaction between task, human scoring, and automated scoring. For instance, although item *sun* showed lower kappa and Pearson correlations as compared to item *spoon*, its kappa and correlation were closer to human/human agreement than those of *spoon*, which points to the need for using multiple criteria in evaluating automated scoring.

Findings showed that length of the responses did not predict scores. The correlation between human rater scores and number of words in the responses was very small, ranging from .09 to .16 for the two human raters. Results for c-rater were similar, ranging from .07 to .14.

Table 6 shows the Pearson correlation between c-rater scores and the original human scores using the holistic rubric, with the values ranging from .67 to .72.

Discussion

Concept-based automated scoring was used to score four science explanation items designed to measure students' understanding of energy-related phenomena such as photosynthesis, energy transfer, and energy transformation. Although four sounds like a small number of items, in the context of automated scoring, testing these items took a substantial amount of effort and spanned 8 months. Using both kappa and Pearson correlation, the consistency between c-rater and human scoring for the items showed moderate to good agreement. The discrepancy between c-rater and human scores varied depending on which criterion was used for evaluation. In the following sections, we discuss the challenges of transforming holistic rubrics to analytic rubrics for automated scoring and the potential benefits of analytic rubrics, addressing Research Question 1. We also comment on the challenges in automated scoring and potential sources of disagreement, addressing Research Questions 2 and 3.

Challenges and Promises of the Analytic Rubrics

Three issues stand out in the transformation of the holistic rubrics to the analytic rubrics. The first one concerns maintaining a balance between adequately preserving the main ideas in the holistic rubric and keeping the number of concepts at a reasonable level. Our initial effort resulted in a large number of concepts (e.g., 12–18) for each item, which made the scoring too cumbersome for human raters to produce reliable scores. The analytic rubric was later consolidated to capture the most important concepts emphasized by the holistic rubric. Human rater agreement improved substantially as a result. Even so, there was still a great deal of variability in terms of human agreement across the four items. For example, the human agreement in kappa was lowest on the *sun* item. Although the analytic rubric only has five concepts for this item, the open nature of the item and the large number of alternative expressions within each concept made it difficult to score. One rater reflected that:

> It's challenging to make a decision about all the possible alternative expressions on this item. There are also many key terms in the model answer (sun, plants, photosynthesis, glucose, etc) and students don't necessarily mention them in a coherent way. The holistic rubrics allow for coding of partial understanding of ideas spread out in different sentences, but the analytic rubric only allows scoring of one sentence independent of others.

The *heat* item had 10 concepts in the analytic rubric and the agreement between the two human raters was .71 in kappa. Although the accuracy was much higher than that of the *sun* item, it was considerably lower than the human/human agreement when they used the holistic rubric (>.90; Liu et al., 2011). The low human agreement using the analytic rubric was possibly a result of the human raters adapting to the new scoring rubric and the new scoring platform. The more concepts, the more difficult it became for human raters to distinguish among the choices.

The second issue concerns the identification of distinct concepts. Some of the concepts may be clearly distinctive to human raters, but their overlapping linguistic features make them difficult for c-rater to differentiate. For example, for the *coal* item ("What possible effect could the increased amount of carbon dioxide have on our planet?"), one of the valid concepts is that "carbon dioxide makes a warmer climate," while an alternative concept is that "carbon dioxide is warm." c-Rater had difficulties distinguishing between these two. For example, one student wrote "It would warm up the planet because carbon dioxide is a green house gas." Human raters can decide that this response captured the valid concept, but c-rater had problems with determining the validity of the response as it contains both "warm" and "carbon dioxide" which are each categorized as an alternative concept.

The third issue lies in the differentiation of valid and invalid ideas coexisting in a response. When using the holistic rubric, raters tend to look for evidence of valid ideas in a holistic way. When valid and invalid ideas coexist in a response, it creates another layer of complexity for holistic human scoring in that the human raters need to make a decision on whether or not to penalize the invalid idea(s), and if so, to what degree (i.e., how to assign scoring weights). In scoring using an analytic rubric, the main ideas are made explicit in the concepts and the scoring rule also specifies the scores based on the combination of valid and invalid concepts. For these reasons, it becomes more straightforward to take both right and wrong ideas under consideration. The following is a sample response to the *spoon* item:

> The metal spoon will feel the hottest because metal can heat up [Concept 4; appendix], and cool down faster than a wooden or a plastic spoon. Also because metal can absorb the heat at a quicker pace [Concept 6].

Using the holistic rubric, human raters assigned a score of 3 (partial-link level) to this response, because it showed partial understanding in the first sentence by mentioning that heat transfers faster in metal than in plastic and wooden spoons. However, the second sentence included an invalid idea by mentioning that the metal spoon "absorbs" heat. This invalid idea was not considered in the holistic scoring rubric. Using the analytic rubric, the first sentence was categorized as a variant of Concept 4 and the second sentence as Concept 6 (appendix). According to the scoring rule, the combination

of Concepts 4 and 6 would yield a score of 2, which is at the no-link level.

The last example illustrates a potential benefit of using the analytic rubric in that it could capture both valid and invalid ideas in the same response and provide nuanced information about student mastery of specific concepts. A holistic score of 3 on an item does not really tell the teacher what the students know and do not know, but a score of 3 with details on each concept will likely provide richer information for the teacher's use. It makes it possible for teachers to analyze student understanding of individual concepts. The analytic rubrics may be of particular value when the assessments are used for diagnostic purposes, so that the teacher can provide targeted feedback to students. The benefits of concept-based scoring lie within the richness of the information it provides and its diagnostic value. Scores produced based on a specific concept, if accurate, can point directly to a student's strengths and weaknesses in understanding that particular concept, which is not something a holistic scoring rubric captures.

An issue that is separate from the typical psychometric evaluation of automated scoring, but is of educational interest, is how well the automated scores are aligned with the scores generated using the original holistic rubrics. From a psychometric evaluation perspective, these two scores are not comparable because they are based on different rubrics. However, from an educational perspective, it is important to know if the automated scores can reflect the same levels of knowledge and understanding emphasized by the holistic rubric. Table 6 shows that the correlations between these two sets of scores are moderate (i.e., .67–.72). However, we suspect that the correlation is a function of the accuracy of the automated scores, because the two items (i.e., *sun and coal*, Table 6 and Figure 1) with lower agreement showed lower correlations. As the levels of accuracy improve, the correlation may increase between these two sets of scores. This will in turn provide further quantitative evidence for the transformation of the holistic rubric into the analytic rubric, in addition to the qualitative evidence gathered through the development process of the analytic rubrics discussed above.

Challenges and Sources of Disagreement in Concept-Based Scoring

This research revealed a number of challenges for the automated scoring of science explanation items. First, it was challenging for c-rater to identify the underlying concepts and capture the places where students' arguments were *incomplete* or *inaccurate*. When middle school students provide explanations to science items, they use vocabulary in unexpected ways. For example, to answer the question "Explain how the sun helps the animals survive," one student wrote "The sun shines down on the plants. The plants turn the sunlight into food. Some animals bite the plant and obtain vitamins and glucose for the animal to live." The human raters and c-rater disagreed on this response because c-rater did not recognize the word *bite* because it was not initially included as a synonym for *eat*. Human raters, however, can exercise their cognitive judgment to decide on the contextual meaning of this word.

Another challenge was to distinguish the nonnormative, incorrect ideas from the normative, correct scientific ideas. c-Rater is only designed to capture the presence or absence of a specific concept. There were a significant number of nonnormative, incorrect ideas in student responses. It would be fairly easy for human raters to evaluate the relevance and correctness of the responses, but this would be challenging for c-rater because the irrelevant responses do not necessarily arise from a coherent conceptual framework that can be captured by the concept scheme used by c-rater. Given the lack of coherency in nonnormative ideas, the only way c-rater could evaluate such responses was to include an individual concept for each nonnormative idea. An obvious limitation of this approach is that it is very difficult, if not impossible, to exhaustively include all nonnormative ideas.

An additional challenge for c-rater and human raters is pronoun resolution. For example, to the same question of "Explain how the sun helps the animals survive," one student wrote "It provides plant with a form of energy which plants use to make food then it is changed into food for animals when they eat it." The pronoun *it* was used three times in this sentence, each time referring to a different object. The first *it* probably refers to the sun, the second refers to energy, and the last refers to the plant. Both c-rater and human raters had difficulty resolving pronoun meanings. Although c-rater has pronoun resolution capability and can even distinguish multiple uses of the same pronoun in the same sentence, it may not always be accurate. To alleviate this problem, students could be advised to be specific and clear in their writing. At the same time, NLP scientists at Educational Testing Service (ETS) are also trying to improve c-rater's pronoun solution so that c-rater can correctly identify pronouns when their use does not cause confusion for human raters.

Finally, the relatively small sample size is very likely another factor that accounted for unreliability in c-rater scoring. The number of responses available for c-rater training was small, ranging from 160 to 206 for each of the four items, and the number of responses used for blind evaluation was smaller, ranging from 80 to 103. We suspect that c-rater's agreement with human raters may increase if c-rater is exposed to greater variation in student responses in both model building and blind evaluation. For example, during model building, if more responses are available more variations of the model answers are likely to be detected. As a result, the linguistic features of alternatives can be incorporated in the model to benefit c-rater scoring of responses in the blind evaluation.

Conclusions and Implications

Results from this study suggest that (1) in its current design, c-rater scores cannot replace human scores, and (2) concept-based automated scoring showed some potential in scoring explanation items with complex scoring rubrics and could serve as a complement to teacher scoring in a low-stakes classroom setting.

The research identified a number of ways to refine the accuracy of c-rater scoring: (1) provide sufficient training to human raters to ensure the reliability of the human scores which serve as a basis for comparison with machine scores, (2) improve automated scoring functionality in terms of dealing with pronouns, (3) develop strategies for defining specific concepts that could offer value in teacher guidance and feedback, and (4) increase the sample size of the responses. Although this study used science explanation items for scoring, the experiences and lessons revealed from this study can be extended to concept-based scoring of items in other domains.

Although concept-based scoring is not yet accurate enough to justify its use as the sole grader on high-stakes tests, it may add value to human scoring of constructed responses for diagnosis and guidance. Prior research shows that effective feedback can be a useful tool to prompt students' reconsideration and refinement of their responses (e.g., Azvedo & Brenard, 1995; Butler & Winne, 1995; Hattie & Temperley, 2007; Meyer et al., 2010; Shute, 2008). When scoring constructed-response items for large classes, very few teachers are able to provide elaborated, specific feedback to each student in the class in a timely manner (Matthews, Janicki, He, & Patterson, 2012; National Council of Teachers of English, 2008). We pilot-tested the effect of automated feedback with two teachers using the *sun* item (Linn & Liu, 2013). Among the 258 students who participated in the pilot, 126 were in the teacher condition where they received feedback from the teacher on the next day, and 132 were in the c-rater condition where they received immediate, automated feedback. Results showed that the students in the c-rater condition were as likely to revisit and revise their responses (85%) as those in the teacher condition (87%). More importantly, students in both conditions made significant and comparable gains through revising their answers (teacher condition, effect size $= .41SD, p < .001$; c-rater condition, effect size $= .38SD, p < .001$). The pilot study provides evidence that automated scoring technology may be valuable for technology-enhanced instruction that features embedded constructed-response items and in which assessment, scoring, and feedback all serve as formative learning opportunities. While automated scores and feedback cannot replace teachers' work, they could serve as a useful complement to teacher scoring so teachers can focus more on lesson planning and helping students in need.

Appendix: c-Rater Concepts, Paraphrases, and Scoring Rule for the *Spoon* Item

Concept 1: 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.

Concept 2: The metal gets hot fastest OR heat will come to it fastest OR metal conducts heat fastest OR metal is the fastest conductor OR heat enters metal faster OR metal absorbs heat faster OR metal will absorb the most heat in the short amount of time OR the metal gets hot in the smallest amount of time OR heat will come to it in the smallest amount of time OR metal conducts heat in the smallest amount of time OR heat enters metal in the smallest amount of time OR metal absorbs heat in the smallest amount of time OR metal transfers heat faster OR metal transfers heat fastest OR metal transfers heat in a smallest amount of time.

Concept 3: Metal conducts the most heat OR more heat comes to the metal OR metal absorbs the most heat OR metal conducts heat more easily OR metal absorbs heat more easily OR metal conducts heat best OR metal is the better conductor OR metal conducts heat better OR metal absorbs heat better OR metal is a greater heat conductor OR metal becomes the hottest OR metal is a good conductor and the others are not OR metal transfers heat "best" OR metal transfers heat more easily OR metal transfers the most heat OR more heat energy is transferred.

Concept 4: Metal heats up OR metal becomes hot OR metal gets hot OR metal gets hot easily OR heat will come to the metal OR metal absorbs heat OR metal absorbs heat easily OR metal conducts heat well OR metal is a conductor OR metal is a very good conductor OR metal is a great conductor OR the heat spreads through the object OR metal transfers heat OR heat enters the metal fast OR heat spreads through the object quick.

Concept 5: Metal attracts heat OR metal attracts the most heat OR metal attracts more heat.

Concept 6: Heat stays in the spoon longer OR metal keeps the heat for the longest time OR heat stays in the spoon longer OR the metal conserves heat OR heat is more apparent in a metal object OR metal feels hotter than wood when it is being heated OR metal extracts heat OR metal absorbs heat OR heat will come to it OR the spoons feel the same OR metal is a good conductor of electricity OR metal conducts cold.

Scoring Rule

4 points C1 and (C2 or C3 or C4)
3 points {C1 and (C2 or C3 or C4)} and C5
3 points (C1 or C2 or C3)
2 points (C1 or C2 or C3) and C5
2 points C4
1 point C4 and C5
1 point C5 or C6 or none

References

Attali, Y., & Powers, D. (2008). *Effect of immediate feedback and revision on psychometric properties of open-ended GRE® subject test items*. GRE Board Research Rep. No. 04-05; ETS RR-08-21. Princeton, NJ: Educational Testing Service.

Attali, Y., Powers, D., Freedman, M., Harrison, M., & Obetz, S. (2008). *Automated scoring of short-answer open-ended GRE subject test items*.ETS GRE Board Research Report No. 04-02. Princeton, NJ: Educational Testing Service.

Azevedo, R., & Brenard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research*, *13*, 11–127.

Bennett, R. E., & Sebrechts, M. M. (1996). The accuracy of expert-system diagnoses of mathematical problem solutions. *Applied Measurement in Education*, *9*, 133–150.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*, 355–377.

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.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Advances in Natural Language Processing*, *18*(1), 32–39.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*, 245–281.

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

Dzikovska, M.O., Moore, J. D., Steinhauser, N., Campbell, G., Farrow, E., & Calloway, C. (2010). Beetle II: A system for tutoring and computational linguistics experimentation. In J. Hajic, S. Carberry, & S. Clark (Eds.), *ACL 2010: Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 13–18). Stroudsburg, PA: Association for Computational Linguistics.

Dzikovska, M. O., Nielsen, R. D., & Brew, C. (2012). Towards effective tutorial feedback for explanation questions: A dataset and baselines. In T. Chandra (Ed.), *Proceedings of the 2012 conference of the North*

American Chapter of the Association for Computational Linguistics: Human language technologies (pp. 200–201). Stroudsburg, PA: 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/

Graesser, A. C. (2011). Learning, thinking, and emoting with discourse teachers technologies. *American Psychologist*, *66*, 743–757.

Graesser, A. C., Rus, V., D'Mello, S. K., & Jackson, G. T. (2008). AutoTutor: Learning through natural language dialogue that adapts to the cognitive and affective states of the learner. In D. H. Robinson & G. Schraw (Eds.), *Recent innovations in educational technology that facilitate student learning* (pp. 95–125). Charlotte, NC: Information Age.

Hattie, J., & Temperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112.

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.

Kintsch, E., Steinhart, D., Stahl, G., Matthews, C., & Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments*, *8*, 87–109.

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

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice*, *23*(3), 6–14.

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. (2010). Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective. *Science Education*, *94*, 665–688.

Lee, H. S., Varma, K., Linn, M. C., & Liu, O. L. (2010). Impact of visualization-based inquiry science experience on classroom learning. *Journal of Research in Science Teaching*, *47*, 71–90.

Linn, M. C., & Eylon, B.-S. (2011). *Science learning and instruction: Taking advantage of technology to promote knowledge integration*. New York, NY: Routledge.

Linn, M. C., & Liu, O. L. (2013, April). NCME invited session on big data: New opportunities for measurement and data analysis. Paper accepted for presentation at the 2013 conference of the National Council on Measurement in Education, San Francisco, CA.

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. (2010). Evaluating inquiry-based science modules using a hierarchical linear model. *Educational Assessment*, *15*(2), 69–86.

Liu, O. L., Lee, H. S., & Linn, M. C. (2011). A comparison among multiple-choice, constructed-response and explanation multiple-choice items. *Educational Assessment*, *16*, 164–184.

Matthews, K., Janicki, T., He, L., & Patterson, L. (2012). Implementation of an automated grading system with an adaptive learning component to affect student feedback and response time. *Journal of Information Systems Education*, *23*(1), 71–83.

Meyer, B. J. F., Wijekumar, K., Middlemiss, W., Higley, K., Lei, P., Meier, C., & Spielvogel, J. (2010). Web-based tutoring of the structure strategy with or without elaborated feedback or choice for fifth- and seventh-grade readers. *Reading Research Quarterly*, *45*, 62–92.

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

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

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

National Council of Teachers of English. (2008). Statement on class size and teacher workload: Secondary. Retrieved November 14, 2013, from http://www.ncte.org/positions/statements/classsizesecondary

Pulman, S., & Sukkarieh, J. (2005). Automatic short answer marking. In *The second workshop on building educational applications using NLP: Proceedings* (pp. 9–16). New Brunswick, NJ: Association for Computational Linguistics. Retrieved November 14, 2013, from http://www.comlab.oxford.ac.uk/people/publications/date/Stephen.Pulman.html

Quellmalz, E., & Pelligrino, J. (2009). Technology and learning. *Science*, *323*, 75–79.

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.

Sargeant, J., Wood, M. M., & Anderson, S. M. (2004, July). A human-computer collaborative approach to the marking of free text answers. Paper presented at the 8th Annual Computer-Assisted Assessment Conference, (pp. 361–370), Loughborough, UK.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*, 153–189.

Sukkarieh, J. Z., & Blackmore, J. (2009, May). 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: AAAI Press.

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

Sukkarieh, J. Z., & Pulman, S. G. (2005). Information extraction and machine learning: Auto-marking short free-text responses for science questions. In *Proceedings of the Twelfth International Conference on Artificial Intelligence in Education* (pp. 629–637). Amsterdam, The Netherlands: IOS Press.

Sukkarieh, J. Z., Pulman, S. G., & Raikes, N. (2003, October). Auto-marking: Using computational linguistics to score short, free text responses. Paper presented at the Twenty-Ninth Annual Conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, *31*, 3–62.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*, 103–118.

Wang, H.-C., Chang, C.-Y., & Li, T.-Y. (2005). Automated scoring for creative problem-solving ability with ideation-explanation modeling. In *Proceedings of the Thirteenth International Conference on Computers in Education* (pp. 522–529). Singapore: IOS Press.

Wiemer-Hastings, P., Arnott, E., & Allbritton, D. (2005). Initial results and mixed directions for research methods tutor. In *Supplementary Proceedings of the Twelfth International Conference on Artificial Intelligence in Education*. Amsterdam, The Netherlands: IOS Press.

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.