Computers and the Humanities **37:** 389–405, 2003.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

C-rater: Automated Scoring of Short-Answer Questions

CLAUDIA LEACOCK[1] and MARTIN CHODOROW[2]
[1]*Educational Testing Service, Rosedale Road, 18-E, Princeton, NJ 08541, USA*
E-mail: cleacock@ets.org
[2]*Hunter College, City University of New York, USA*
E-mail: martin.chodorow@hunter.cuny.edu

Abstract. C-rater is an automated scoring engine that has been developed to score responses to content-based short answer questions. It is *not* simply a string matching program – instead it uses predicate argument structure, pronominal reference, morphological analysis and synonyms to assign full or partial credit to a short answer question. C-rater has been used in two studies: National Assessment for Educational Progress (NAEP) and a statewide assessment in Indiana. In both studies, c-rater agreed with human graders about 84% of the time.

Key words: automated scoring, content-based scoring, short answer scoring

1. Introduction

As more and more assessment, test preparation and instructional materials are delivered online, the possibility for scoring test items automatically becomes a reality and so too does the need. Vigilante reports that in the New York University Virtual College "25 percent of online faculty time [is] currently spent on grading written assignments and examinations" (Vigilante, 1999, p. 59). To date, research in automated scoring has focused on scoring essay-length responses as opposed to short content-based answers that are typically written for homework assignments, classroom tests, and end-of-chapter review questions. The scope of these short-answer questions can range from mathematics, where students are asked to explain how they arrived at their answer, to science, where they may be required to describe an experiment or define a term, to reading comprehension questions, or to history, where they may have to describe or discuss an event (see Table I for examples). Of course, regardless of the subject domain the goal of automated scoring is to provide immediate and accurate feedback to the student.

C-rater^{TM} is a short-answer scoring engine, developed by ETS Technologies, which is designed for constructed-response answers to questions that measure understanding of content materials. C-rater differs from essay scoring systems (Burstein, 2003; Eliot, 2003; Landauer, 2003) in several fundamental and important ways that reflect its primary task – recognizing paraphrase or equivalent meaning.

Table I. Example questions that have been scored by c-rater

Grade	Subject	Question
8	Science	Explain how you would design an experiment that would investigate the importance of light to plant growth. Include the type of organisms required, the control and variable, and the method of measuring results.
8	Math	A radio station wanted to determine the most popular type of music among those in the listening range of the station. Would sampling opinions at a Country Music Concert held in the listening area of the station be a good way to do this? Explain your answer.
11	Reading comprehension	Compare and contrast what Mama and Walter in *A Raisin in the Sun* believe to be the most important thing in life or what they "dream" of. Support your choice for each character with dialogue from the excerpt of the play.
College	Database management	Differentiate between logical and physical models.

It begins with a model of the correct answer that is created by a content expert. C-rater's goal is to map the student's response onto the model, and in so doing to demonstrate the correctness of the response or, failing that, its incorrectness or inadequacy. The model is constructed by hand but the mapping is fully automated. Because a model is required, the question must have a single correct answer or a range of correct answers. This means that C-rater is *not* designed to score open-ended questions, such as ones that ask for examples taken from personal experience, or for an opinion, or for innovative approaches to resolving a conflict. But it can score questions that look for specific ideas, such as those from science, math, reading comprehension, and database management shown in Table I.

What is involved in scoring the response to a content-based question? A question is designed to elicit from the student one or more concepts that constitute the correct answer. However, there are an enormous number of ways that a single concept can be expressed in natural language. To score short answer responses, the scoring engine must be able to recognize when a concept is expressed and when it is not. We think of the set of correct responses as being paraphrases of the correct answer, and of the c-rater scoring engine as a *paraphrase recognizer* that identifies the members of this set.

C-rater analyzes responses using a range of natural language processing techniques. It normalizes across the variety of ways a single concept can be expressed by focusing on four primary sources of variation among sentences: syntactic variation such as an active sentence versus a passive one ("*You need two plants*" and

"*Two plants are needed*"); pronoun reference ("*Mama believes that dignity is important.* ***She*** *said* ...); morphological variation (*believed, believing, beliefs*); and the use of synonyms and similar words (*subtract, minus, less than*). The system also handles a fifth source of variation – the variation caused by typographical and spelling errors. Although spelling is not a source of variation that is usually considered when studying paraphrases, recognizing which word the student intended to type is critical for a scoring engine. Once a student response has been normalized into a *canonical representation*, c-rater tries to match the concepts it has identified to the concepts that are represented in the model of the correct answer. It then assigns a score depending on the number of concepts that are matched.

C-rater is *not* a simple word-matching engine that looks for a list of key words and assigns a score without regard to their context. Instead, the concepts that c-rater identifies are typically sentential rather than atomic.

As noted earlier, much of the time spent by the faculty in online courses is devoted to grading written assignments. The same is true for the classroom teacher. C-rater was developed to save classroom time and to give students immediate feedback. If a teacher uses a short-answer question just once for a single class, then there is no reason to devote time generating a c-rater model. However, if the teacher uses the same question for several classes or over several semesters, then the advantages of the initial effort are worthwhile. In large-scale assessments, the advantage of using automated scoring is even more marked. For example, during a six-week window in Spring 2002, c-rater was used to score about 100,000 responses to 11th grade reading comprehension questions for the state of Indiana.

2. How C-rater Works

Paraphrases of a concept, even a very specific concept, can vary enormously. Consider the following responses that describe the experimental design in the eighth grade science question from Table I:

- Put one plant under the light and another in a very dark closet.
- Keep one plant in the sun, the other in the dark.
- I would put one in the sunlight and one where there is no light.
- You would need one plant near the light as the control and another away from the light as a variable.

All of these sentences convey the same idea, even though they differ in a number of ways. The first two are in the form of imperatives while the second two contain subjects. *Plants* are variously referred to as *another, the other*, and *one*. The concept of *darkness* can be expressed as *no light* and *away from the light. Light* is expressed as *sunlight* or *sun*. Although this is not an exhaustive list of the differences between these sentences, it becomes apparent that the differences are not trivial. Compare these to the following incorrect response:

- A *plant* will live in the *light* than *dark*

It has vocabulary very similar to the paraphrases of the correct answer, yet judges marked it as being incorrect. The challenge for automated scoring is to determine that the first four sentences are paraphrases while the fifth is not.

In order to recognize that responses express a common meaning, c-rater generates a canonical representation of each response. To build this representation, it extracts the underlying structure of the response, resolves pronoun reference, normalizes across inflected words, and recognizes the use of similar terms and synonyms. The next sections illustrate the architecture of c-rater and how it builds a canonical representation of a response. The purpose is to provide an understanding of its operation rather than to give a detailed technical description.

2.1. Spelling correction in a restricted domain

In word processors, spelling correction is almost always interactive. When a word is not found in the dictionary, a menu of possible words is displayed in a pop-up window and the user can select the word that was intended. Spelling correction is typically interactive because the spell checker does not know the semantic domain of the text. For example, when faced with "Reagons", it is perfectly reasonable for a word processor to suggest, as a first choice, the noun "Reasons". However, if the domain of discourse is about recent US presidents, the suggestion will, more likely than not, be wrong.

For responses to content-based questions however, the semantic domain is highly restricted, consisting of the language in the question, the reading passages, and the model answer. This restricted semantic domain enables c-rater to perform accurate, behind-the-scenes, automatic spelling correction. As an example, one of the questions that c-rater has scored asked about challenges facing incoming presidents. One correct response was "Ragen addressed the need to end the bout of inflation that plagued the nation". For c-rater to recognize that this student response is correct, it must replace *Ragen* with *Reagan*. To give some sense of the magnitude of this problem, we discovered 67 different variants of *Reagan* in about 9,000 responses. Below are all the spelling variants of *Reagan* that occurred more than once:

> Regan, Reagon, Reagen, Raegan, Regans, Regean, Reagons, Ragan, Ragen, Reagin, Raegon, Regon, Reagn, Reagean, Reegan, Ragon, Ragean, Reagens, Raegen, Raegans, Reggan, Raygon, Rgan, Regens, Regen, Regeans, Reagion, Ragons, Raegin

C-rater's spelling correction module recognizes a misspelled word when the morphological analyzer cannot find a base form in its dictionary (Fellbaum, 1998). It then uses an edit distance algorithm (Cormen *et al.*, 2001) to compute the number of keystrokes that separate the unrecognized word from the words in the semantic domain of the question. When the minimum edit distance is small, the unrecognized word is replaced with the closest word in the question's semantic domain. In this way, 84% of the variants of *Reagan/Reagan's* were correctly identified.

Table II. Tuples for 4 responses

Score	Sentence and tuple		
Credit	Most people at the country show would say that country music is the most popular music.		
	say	**:subject most people**	
	be	**:subject country music**	**:object most popular music**
Credit	The people at the country concert would all answer country music.		
	answer	**:subject people**	**:object country music**
Credit	People at a country concert might think that country music is the best music.		
	think	**:subject people**	
	be	**:subject country music**	**:object best music**
No credit	I happen to like country music and so do most of my friends.		
	like	**:subject I**	**:object country music**
	do	**:object most of my friends**	

Of course, not all typographical or spelling errors result in nonwords. C-rater is unable to detect mistakes such as “add umber” and “ode nuber” (both of which appeared in 4th grade responses) when the student meant “odd number”, because “add”, “ode”, and “umber” are all English words.

2.2. SYNTACTIC VARIETY

Much of the variation in responses is due to differences in surface syntax. To recover a canonical syntactic form c-rater first generates a shallow syntactic analysis (Abney, 1996) from which it extracts the predicate argument structure, or *tuples*, of each sentence in the response. A tuple consist of the verb in each clause along with its arguments (such as subject and object) and complements (such prepositional phrases). Table II shows the relevant elements of the tuples for three correct responses and one incorrect response to the eighth grade question about sampling at a country music concert. Although the surface structures of the three correct responses are quite different, their underlying structures are similar. The subject of the main clause is “people”, and the object of either the main clause or a subordinate clause is “music”.

If one ignores the predicate argument structure of the responses and looks only at the language of the response without regard to word order, incorrect responses are likely to receive credit. For example, the “no credit” response in Table II shares much of the language of the correct responses, yet it was judged to be incorrect by the human readers. Notice that, although the language is similar, the tuples are not. In this sentence the subject is "I", not the people at the concert who are being interviewed.

Once the tuples have been generated, c-rater stops working with the original sentences and normalizes the tuples instead, thereby eliminating many of the surface differences that appear in the paraphrases.

It is a matter of debate whether contextual information is required for scoring essay-length passages for content. Landauer *et al.* (1997) find that contextual information is not important when their latent semantic analysis (LSA) system scores essays for content:

> The fact that LSA can capture as much of meaning as it does without using word order shows that the mere combination of words in passages constrains overall meaning very strongly.

Systems, such as LSA, that do not use contextual information are called "bag-of-word" approaches because they treat a response as simply that – a set of unordered words. We have found that word order *is* important, at least for scoring short answer responses, a point that we will return to later.

2.3. PRONOUN RESOLUTION

After the response has been represented as predicate-argument tuples, the next step is to identify the referents of any pronouns it contains. The pronoun resolution component is a version of Morton (2000) that has been specifically trained on student responses to essays and short-answer questions. It identifies all of the noun phrases that precede the pronoun, as well as all of the noun phrases in the question, and selects the one which the pronoun is most likely to refer to.

Pronoun resolution proved to be particularly important in a question that asked students to read passages and identify the issues that three U.S. Presidents emphasized. Since the presidents were all male, the pronoun "he" gave no clue as to which president's ideas were being discussed. More typically, the pronoun that needs to be resolved in student responses is "it" as in "Take one plant and set it in a dark closet":

set :object it :in dark closet
pronoun resolution module
set :object one plant :in dark closet

2.4. MORPHOLOGY

Next, c-rater then normalizes across variations in word form – substituting the base form for each inflected word in the tuple. The morphological analysis component recognizes two kinds of morphological variation: *inflectional* and *derivational.*

Inflectional morphology consists of those grammatical markers that attach to words in order to indicate, for instance, plurality in nouns and tense in verbs. For example, *subtracts*, *subtracting* and *subtracted* share the same base form: *subtract*. Derivational morphology involves a change in the syntactic category of a word. The attachment of a suffix results in the derivation of a new part of speech, as shown by the difference between the verb *subtract*s and the noun *subtraction*, where the suffix *-tion* has been added to the verb to derive a nominal form. However, the underlying stem, *subtract*, is the same for both the noun and the verb.

I used subtraction.
use :subject I :object subtract
I subtracted 5.
subtract :subject I :object 5

2.4.1. *Morphology and Negation*

Negating prefixes, such as *un-* are also stripped from words, but their meaning is retained as *not* in the tuple (see below).

a. The sample is unfair.
   be fair :not :subject sample

b. The sample is not fair.
   be fair :not :subject sample

This makes morphological negation equivalent to lexical negation. Incorporating negation into the tuple solves a problem that the bag-of-words approach cannot handle – it provides a way to mark the scope of negation. If the sentence "a plant will only live in the light and not the dark" is represented as an unordered bag of words, it is no longer possible to distinguish it from "a plant will only live in the dark and not the light" or "a plant will not only in the light and the dark". In the tuple, the proper association between *live, not* and *dark* can be maintained.

2.5. Filling in the Semantic Gaps

The final step is c-rater's lexical substitution to normalize for word meaning. C-rater uses for this purpose a statistically generated *word similarity matrix* (Lin, 1998) that was trained on more than 300 million words of current American and British fiction, nonfiction and textbooks. The matrix was generated by a program that produces a shallow parse of text and then computes word similarities based on the overlap of the words' contexts. The underlying idea is that words that appear in the same contexts are likely to be similar to one another. Intuitively, if one inspects all of the nouns that are objects of "cook" in a 300-million word corpus, one will find a long list of foods that get cooked. Using this approach, the program found,

Table III. Word similarity matrices

Headword	Similar terms
choose	select elect decide nominate pick appoint adopt designate prefer want approve vote for determine endorse prepare consider favor accept reelect hire *reject* vote mention recommend propose discuss . . .
biased	misleading erroneous discriminatory one-sided slanderous *unbiased* inaccurate prejudiced incorrect distorted unfair irresponsible subjective racist untrue unfounded coercive skewed inequitable false incomplete . . .

for example, that the verbs *select* and *choose* often appear with the same objects and subjects and are therefore likely to be similar. Table III shows a portion of the entries in the matrix for *choose* and *biased.* Words in each entry are listed in decreasing order of similarity. According to this measure, *misleading* is the most similar word to *biased,* while *incomplete* is less similar. It is important to note that the word similarity matrix does not list synonyms as such: *biased* and *misleading* are not synonyms, but they are similar in that they often appear in similar contexts.

As it happens, antonyms are also often used in similar contexts. In our word similarity matrix, the adjective that is most similar to “good” is “bad.” In Table III, *reject* is similar to *choose* and *unbiased* is similar to *biased.* This problem is not exclusive to Lin’s approach but is common to any statistical method for finding similar words based on similar contexts. Because of this, when a content expert creates the model answer, the process described in the next section, the expert is given an opportunity to remove from the similar words list any that are antonyms or are otherwise inappropriate.

As a response is evaluated, each base form in the response is checked against the base forms in the model answer and their synonym/similar word lists. Once a match between the response and the base form lists is found, the word in the response is replaced with the word from the model answer.

2.6. CONCEPT MATCHING

After the canonical representation of the response is completed, the final step is to compare it to the canonical representation of the model answer. The algorithm that matches student responses to the model answer is rule-based. For example, one rule requires that, in the absence of a passive construction, subjects and objects cannot be interchanged (except for a small class of verbs). This prevents “the man bit a dog” and “a dog bit a man” from being recognized as paraphrases but allows “the man was bitten by the dog”. However, since many of the responses are ungrammatical or fragmentary, the matching algorithm is fairly forgiving. In allowing for various degrees of ungrammatical input, there is a tradeoff. If it is strictly enforced, then too many correct answers will be missed. If it is too lax, then the order problem of the "bag-of-words" approach appears and too many incorrect responses are given credit.

To summarize, c-rater's strategy is to extract and normalize predicates and their arguments. Then, for each relation in the gold standard canonical representation, c-rater tries to find a comparable relation in the response. There will not always be a one-to-one correspondence between arguments in the canonical representation of the model answer and those in the correct responses. A content expert specifies those elements that are required in a response during the process of building model answers to the questions.

3. Building the Model

In an early collaboration with the NYU Virtual College, we derived a model directly from the scoring rubric. Three questions from the chapter review sections of a database management textbook were included the course's final exam. All of the questions were definitional, like the database management question in Table I. In this experiment, the answer provided in the teacher's manual was used as the single model answer for each question and the inclusion of synonyms and similar words was fully automated. While this experiment produced fairly good results (an average of 82.6% agreement with the faculty member who scored the test), it became clear that fully automating the process without intervention by a content expert is not feasible.

Often the concept specified in a rubric, or even in the teacher's edition of a textbook, is not a good match for student responses. Consider the rubric in Table IV, which states that, to receive credit, the response must indicate that the sample is biased. In a total of 1,000 responses, only 16 students used the word *bias*ed (13 of which were spelled correctly) and not many used synonyms of *biased* either. By far, the most frequent correct response to this question was some variation on "People at a country music show would choose country music". It would be quite a stretch for any artificial intelligence system to recognize "People would say they like country music" as a paraphrase of "The sample is biased".

3.1. THE *ALCHEMIST* INTERFACE

An interface called *Alchemist* was designed to guide the content expert through the process of creating the model answers to each question. Its purpose is to provide a bridge between the scoring rubric and acceptable responses that a student is likely to give.

Table V shows the 8th grade science question in Table I that asks the student to design an experiment involving the effects of photosynthesis. According to the rubric, there are four *essential points* that a response must include in order to receive full credit. If a response only addresses one, two, or three of the essential points, partial credit is assigned. If no essential point is covered, no credit is assigned.

Table IV. 8th Grade NAEP math. This is an approximation of a prompt used in the NAEP study

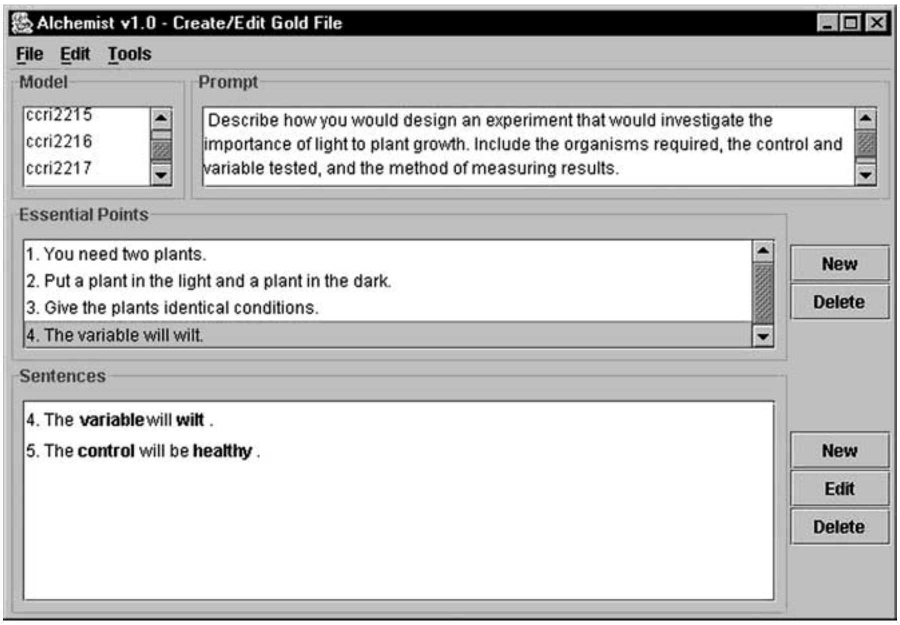
Question	A radio station wanted to determine the most popular type of music among those in the listening range of the station. Would sampling opinions at a Country Music Concert held in the listening area of the station be a good way to do this? ○YES ○NO Explain your answer.
Scoring	Assign full credit if the answer indicates:
rubric	No. The opinions would most likely be biased in favor of those who like country music.

Table V. 8th grade science question

Question	Describe how you would design an experiment that would investigate the importance of light to plant growth. Include the organisms required, the control and variable tested, and the method of measuring results.
Scoring rubric	To receive full credit, response must contain all four of the elements below. To receive partial credit, the response must contain one, two or three of the elements below: 1. The need for two plants; 2. The need for a control grown in light and another plant grown in the dark; 3. The need for all other factors and conditions to be the same, except light, which is the variable condition. 4. The need for some kind of measurement of plant growth or health.

The first step in using the interface is to break down the concepts, so far as possible, into simple sentences. For example, the most frequent sentence that covered the first essential point was: “You need two plants”. This step is greatly facilitated if some scored student responses from are available. In order to generate robust models for the NAEP study and Indiana pilot, the models were built based on the inspection of about 100 scored pretest responses for each question.

Figure 1 shows the *Alchemist* interface for building the model responses. The box on the top left shows the identification number of the question, with the text of the question to the right. Underneath is a listing of the essential points that are specified by the content expert who is creating the model. As can be seen, there are four essential points that need to be identified to get full credit. For each essential element, any number of sentences can be entered, revised or deleted. Figure 1 shows that the fourth point is about measuring the plant’s growth or health. This part of the question is typically answered by noting either that the control will be healthy or that the variable will be unhealthy. The box labeled “Sentences” displays all of the sentences that have been entered for the highlightessential point.



ccri2215
ccri2216
ccri2217

Describe how you would design an experiment that would investigate the importance of light to plant growth. Include the organisms required, the control and variable tested, and the method of measuring results.

1. You need two plants.
2. Put a plant in the light and a plant in the dark.
3. Give the plants identical conditions.
4. The variable will wilt.

4. The **variable** will **wilt** .
5. The **control** will be **healthy** .

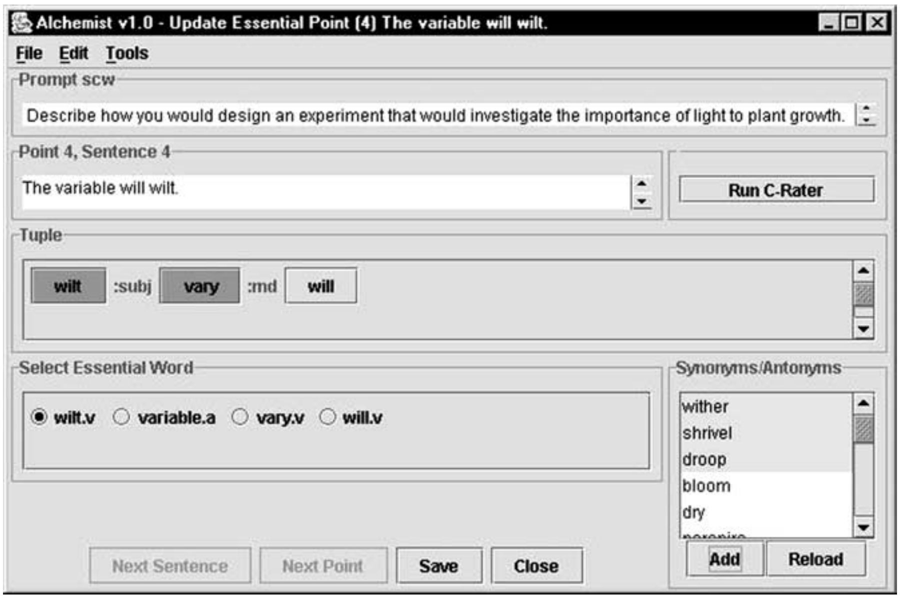
Figure 1. Interface for defining essential points.

As each sentence is entered, c-rater generates a canonical representation for it. This is the representation that c-rater uses. Figure 2 shows how a content expert specifies which atoms in the model answer are crucial to the response by highlighting the words in the tuple box. In this case, a subject noun that is represented by “variable” and a predicate that is represented by “wilt” are required. A list of similar words for each highlighted word is then displayed – from which appropriate words can be selected. In Figure 2, “wither”, “shrivel” and “droop” have been selected as suitable similar words, while “bloom” and “dry” have not. If a synonym is missing from the set, for example, “die” is not similar to “wilt”, but in this context it is an acceptable substitute, words can be added to the set by clicking on the “Add” button. When this process has been completed for each sentence in the model answer, then c-rater is ready to run in its scoring mode.

4. Case Studies

Computer programs that assign a holistic score to an essay have been commercially available for several years. For example, e-rater (Burstein, 2003) has been used to score Graduate Management Admissions Tests since 1999. Other services for assigning an holistic score to essays include those described in Landauer (2002) and in Elliot (2002).

Ever since the Educational Testing Service (ETS) began planning for large-scale computer-based testing, researchers have been developing and evaluating methods for automated scoring of free-response questions (cf. Kaplan and Bennett, 1994;



Alchemist v1.0 - Update Essential Point (4) The variable will wilt.

Prompt scw

Describe how you would design an experiment that would investigate the importance of light to plant growth.

Point 4, Sentence 4

The variable will wilt.

Tuple

wilt :subj vary :md will

Select Essential Word

● wilt.v ○ variable.a ○ vary.v ○ will.v

Synonyms/Antonyms

wither
shrivel
droop
bloom
dry

Figure 2. Specifying key elements of the concept and similar words.

Kud *et al.*, 1994; Burstein *et al.*, 1999). These research projects, however, focused on scoring new question types that were being considered for inclusion in ETS administered tests.

This work on scoring short-answer responses originated with a pilot study in a collaboration between ETS and the NYU Virtual College. Subsequently, c-rater has been used in a study for scoring 4th and 8th grade math questions for the National Assessment of Educational Progress (NAEP) and in a statewide Indiana pilot assessment.

The c-rater engine has been tested and evaluated in two large-scale assessment programs. The first is the NAEP Math Online project, an experimental study exploring the potential uses of technology for the NAEP assessments (Sandine *et al.*, 2002). C-rater evaluated students' written explanations of the reasoning or processes they used to solve math problems. In the second study, which took place in the spring of 2002, c-rater was deployed in the online administration and scoring of Indiana's English 11 End of Course Assessment pilot study. This assessment included seven short-answer questions related to literature selections. C-rater scored over 100,000 11th grade student responses to the reading comprehension questions in this end-of-year test. In these experiments, none of the test questions were designed with c-rater in mind. In fact, those who developed the questions were not even aware of its existence.

The answer models were generated using the Alchemist interface. The model answers were manually generated after inspecting between 60 and 100 scored

Table VI. Percentage of agreement between human readers and c-rater

Grade	Question Number (Point Scale)	Reader 1 = Reader 2 (kappa)	C-rater = Reader 1 (kappa)	C-rater = Reader 2 (kappa)
4	NAEP 1 (3)	94 (0.90)	83 (0.75)	81 (0.71)
8	NAEP 2 (3)	92 (0.86)	91 (0.86)	90 (0.83)
8	NAEP 3 (2)	91 (0.79)	80 (0.58)	81 (0.60)
8	NAEP 4 (3)	90 (0.85)	83 (0.72)	81 (0.69)
8	NAEP 5 (5)	87 (0.77)	85 (0.75)	85 (0.74)

responses to each question. These models were then cross validated using another set of about 100 scored responses that we did not see.

4.1. NAEP MATH

In the NAEP assessment, students were asked to explain how they arrived at their conclusion. The average length of the responses was 1.2 sentences or 15 words. Between 245 and 250 randomly chosen student responses were scored by two human judges and by c-rater. Table II shows the grade level of the question, the number of points on the scoring scale (ranging from two to five score points), the percentage of the time that the two judges agreed with each other, the percentage that c-rater agreed with the first reader and with the second reader. These agreement percentages are accompanied by kappa values, which correct for the level of agreement that is expected by chance. As stated by Fleiss (1981), "Values greater than 0.75 or so may be taken to represent excellent agreement beyond chance, values below 0.40 or so may be taken to represent poor agreement beyond chance, and values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance".

When c-rater was not in agreement with one or both of the judges, the scores were resolved by an expert judge. In all, 35% of the discrepant scores were resolved by the expert in favor of c-rater.

4.2. INDIANA PILOT STUDY: READING COMPREHENSION

In the Indiana Pilot, c-rater was used to score 16,625 reading comprehension responses for each of seven questions. Each response was assigned full credit (2), partial credit (1) or no credit (0). These questions were more open-ended than the mathematics responses, as can be seen from the example in Table I. They were also considerably longer, with an average length of 2.8 sentences or 43 words. In order to evaluate c-rater's accuracy, 100 responses to each question were randomly

Table VII. Human c-rater agreement in the Indiana Pilot

Question Number	Human & C-rater Agreement and Kappas	Baseline
Indiana 1	83% (0.69)	55%
Indiana 2	89% (0.78)	63%
Indiana 3	85% (0.77)	44%
Indiana 4	85% (0.78)	34%
Indiana 5	88% (0.79)	45%
Indiana 6	80% (0.68)	42%
Indiana 7	79% (0.66)	48%
Average	84% (0.74)	47%

sampled and scored by a human judge. When the judge and c-rater assigned different scores, a second reader resolved the score. Overall, c-rater performed with 84% accuracy. The baseline reported in Table VII shows how a system that always assigned the most frequent score would do. In the case of Indiana 1, readers assigned the score of 1 (partial credit) 55% of the time. Therefore, a system that always assigned partial credit to all of the responses to Indiana 1 would be accurate 55% of the time.

C-rater performed very well, at close to 90% accuracy, on two of the questions, while it performed less well, close to 80% accuracy, on two other questions. On average, c-rater and the readers were in agreement 84% of the time. For the two lowest scoring questions, it should be noted that there were very few correct responses in the pretest data – the data that were used to build and cross-validate the model answers. In the case of Indiana 6, there were only seven full-credit responses in the pretest set, and only 15 full-credit responses for Indiana 7.

The combined confusion matrix for all seven questions is shown in Table VIII. The top number in each cell shows the frequency of the responses (for a total of 700), while the percentage is shown below. The first cell in the top row indicates that c-rater and the reader both gave zeros to 229 out of the 700 responses or 32.7% of the time. The third cell in the row indicates that c-rater assigned a zero (no credit) to a correct response (full-credit) 9 times out of 700, or 1.2%. The cells that are shaded show where the computer and the judge were in agreement.

When c-rater made an error, it was usually off by a single score point – 14.4% of the time. In 6.2% of its scores, there was confusion between no credit (score 0) and partial credit (score 1). The remaining 8.2% was confusion between partial credit and full credit. C-rater was confused between no credit and full credit nine out of 700 times (1.2%).

In order to see how effectively a simple bag-of-words approach could score short-answer responses, we scored the Indiana and NAEP responses using a

Table VIII. Confusion matrix for the Indiana Pilot. 0 is no credit, 1 is partial credit, 2 is full credit

	Human 0	Human 1	Human 2
C-rater 0	229 32.7%	25 3.5%	9 1.2%
C-rater 1	19 2.7%	258 36.8%	47 6.7%
C-rater 2	0	11 1.5%	102 14.5%

simple content vector analysis (CVA) classifier based on the vector space model commonly used in information retrieval (Salton, 1975). To return to the "bag of words" analogy, CVA processes a sample of graded responses and, for each score point, it creates a bag of words that appeared in the answers. It then compares the words in every new response to those in each bag by means of a cosine correlation and assigns the score of the bag that is most similar to the response. When we used this "bag-of-words" approach, performance dropped an average of 12% on the NAEP data and 30% on the Indiana Pilot data. Results were further degraded in a variant of this CVA procedure where each new response was compared to every individual training response and the score of the most similar one was assigned. We conclude from these experiments that c-rater's use of predicate argument structure and similar words are responsible for its superior results.

5. Sources of Error: When C-rater Fails

We manually inspected the errors for the Indiana pilot to determine their sources. The errors fall into two categories: *misses* and *false positives*. A *miss* occurs when a response does not get credit for one or more concepts that it, in fact, contains – a response does not get enough credit. A *false positive* occurs when c-rater assigns too much credit to a response – assigning credit for concepts that it does not contain. In the Indiana study, 73% of the errors were misses while only 29% of the NAEP errors were misses. This relatively high ratio of misses in the Indiana pilot as compared to NAEP may reflect the much more open-ended nature of reading comprehension questions as compared to the math questions.

If a correct response is expressed in a truly original manner, c-rater probably won't recognize it and will assign too little credit. In the Indiana Pilot, one of the concepts in the model is that "fire alarms are expensive". When a student responded that "fire alarms take a chunk of change", c-rater did not recognize that "take a chunk of change" means the same thing as "expensive".

There are two reasons for c-rater's false positives. The first is when a student does not know when to stop typing – beginning with a correct answer but going on to say something that is clearly wrong. C-rater is designed to look recognize a correct answer. It assigns credit when it identifies the concepts that it is looking for. It does not *look* for wrong answers. When a student response contains the concept that c-rater is looking for but then goes on and adds something that makes it clear to the readers that he or she does get the point. The main reason, however, for false positives is that the student happens to use the correct language – but that the language is used in such a manner that it does not, in fact, convey the concept. Many of these false positives are the result of allowing for ungrammatical and fragmentary responses.

6. Assessment and Instructional Uses for C-rater

We envision two uses for c-rater: one as an assessment tool and the other as an instructional tool. The NAEP and Indiana experiments were both large-scale assessments while the NYU collaboration was on a classroom scale. In these studies, the responses were scored in batch mode. In the case of Indiana, scored responses were returned within three days. The next step is to implement a web-based version of c-rater where student can get scores immediately. And, in addition to the score, we could show what part of the response received credit and what part of the model correct answer c-rater is unable to identify.

In instructional contexts, c-rater can serve as an adaptive learning tool. As c-rater evaluates a student's understanding of a key concept, it can use the results of the evaluation to direct the student to a location in an online text that contains information that is missing from the student's response, or direct the student to more extensive information on the concept if the student's response shows a targeted level of understanding.

Acknowledgements

We are indebted to Eleanor Bolge for her many contributions to c-rater, John Blackmore for developing the Alchemist interface, and Chi Lu for running the CVA experiment. We also thank Ray C. Dougherty and two anonymous reviewers for helpful suggestions.

Any opinions expressed here are those of the authors and not necessarily of the Educational Testing Service.

References

Abney S. (1996). Partial Parsing via Finite-State Cascades. *Proceedings of the ESSLLI '96 Robust Parsing Workshop.*

Burstein J., Wolff S., Lu C. (1999). Using Lexical Semantic Techniques to Classify Free-responses. In Ide N. and Veronis J. (eds.), *The Depth and Breadth of Semantic Lexicons*, Kluwer Academic Press.

Burstein J. (2003). The E-rater® Scoring Engine: Automated Essay Scoring with Natural Language Processing. In Shermis M.D. and Burstein J. (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum, Mahwah, NJ.

Cormen T.H., Leiserson C.E., Rivest R.L., Stein C. (2001). *Introduction to Algorithms, Second Edition*. The MIT Press, Cambridge, MA.

Elliott S. (2003) Intellemetric: From Here to Validity. In Shermis M.D. and Burstein J. (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum, Mahwah, NJ.

Fellbaum C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge.

Fleiss J.L. (1981). *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York. pp. 212–236.

Landauer T.K., Laham D., Foltz P. (2003). Automated Scoring and Annotation of Essays with Intelligent Essay Assessor™. In Shermis M.D. and Burstein J. (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum, Mahwah, NJ.

Landauer T.K., Laham D., Rehder B., Schreiner M.E. (1997). How Well Can Passage Meaning be Derived Without Using Word Order? A Comparison of Latent Semantic Analysis and Humans. In Shafto M.G. and Langley P. (eds.), *Proceedings of the 19th Anual Meeting of the Cognitive Science Society*, Lawrence Erlbaum, Mahwah, NJ, pp. 412–417.

Kaplan R.M., Bennett R.E. (1994). Using the Free-response Scoring Tool to Automatically Score the Formulating-hypotheses Item. *ETS Research Report 04–08.*

Kud J.M., Krupka G.R., Rau L.F. (1994). Methods for Categorizing Short Answer Responses. *Proceedings of the Educational Testing Service Conference on Natural Language Processing Techniques and Technology in Education and Assessment*. Princeton, NJ.

Lin D. (1998). Automatic Retrieval and Clustering of Similar Words. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistic*, Montreal, pp. 898–904.

Morton T.S. (2000). Coreference for NLP applications. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong.

Salton G., Wong A., Yang C.S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18/11, pp. 613–620.

Sandene B., Bennett R., Braswell J., Oranje A. (Forthcoming). Mathematics Online Study: Final Report. National Center for Education Statistics, Washington, DC.

Vigilante R. (1999). Online Computer Scoring of Constructed-response Questions. *Journal of Information Technology Impact*, 1/2, pp. 57–62.