Listening. Learning. Leading.®

Automated Scoring of Short-Answer Open-Ended GRE Subject Test Items

Yigal Attali

Don Powers

Marshall Freedman

Marissa Harrison

Susan Obetz

April 2008

ETS GRE Board Research Report No. 04-02

ETS RR-08-20

Automated Scoring of Short-Answer Open-Ended GRE® Subject Test Items

Yigal Attali, Don Powers, Marshall Freedman, Marissa Harrison,[1] and Susan Obetz

ETS, Princeton, NJ

GRE Board Research Report No. 04-02

ETS RR-08-20

April 2008

The report presents the findings of a
research project funded by and carried
out under the auspices of the Graduate
Record Examinations Board.

Educational Testing Service, Princeton, NJ 08541

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board reports do no necessarily represent official Graduate Record Examinations Board position or policy.

The Graduate Record Examinations and ETS are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

****

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

Educational Testing Service
Princeton, NJ 08541

Abstract

This report describes the development, administration, and scoring of open-ended variants of GRE® Subject Test items in biology and psychology. These questions were administered in a Web-based experiment to registered examinees of the respective Subject Tests. The questions required a short answer of 1-3 sentences, and responses were automatically scored by natural language processing methods, using the *c-rater*™ scoring engine, immediately after participants submitted their responses. Participants received immediate feedback on the correctness of their answers, and an opportunity to revise their answers. Subsequent human scoring of the responses allowed an evaluation of the quality of automated scoring. This report focuses on the success of the automated scoring process. A separate report describes the feedback and revision results.

Key words: Automated scoring, constructed response, c-rater, GRE Subject test

Introduction

In contrast to multiple-choice (MC) questions, constructed-response (CR) questions require students to write their own answer and allow students to express and support their ideas in response to the text. By doing so, students can demonstrate a range of abilities: describing interpretations, explaining personal reactions, generating conclusions, or supporting critical evaluations. However, the scoring of these CR items is expensive, time-consuming, and adds potential measurement error to test results due to inconsistencies in the scoring process.

This study examined the possibility of automatically scoring short-answer CR questions for the GRE® Subject Tests, which are currently composed solely of MC questions. These tests assess knowledge of a specific subject area acquired over time in many undergraduate courses. Validity studies have consistently shown GRE Subject Test scores to be better predictors of success in graduate school than are GRE General Test scores (see, for example, Briel, O'Neill, & Scheuneman, 1993; Kuncel, Hezlett, & Ones, 2000).

On the other hand, the Subject Tests have sometimes been criticized for measuring only superficial knowledge of a discipline, rather than a deep understanding of subject area (e.g., Glanz, 1996). The question in Figure 1, which appears in the *GRE Psychology Test Practice Book* (2001), calls for an explanation of a phenomenon. However, in the context of an MC question the recognition task is strongly influenced by the alternatives and the relations among them. Consequently, performance may be influenced by general knowledge or reasoning skills. An alternative CR task could ask the student to explain this phenomenon without the explicit guidance of alternatives, and thus could be more informative about the student's knowledge of the subject area.

Scoring these CR alternatives to subject area questions is a challenge even for human raters. In this study we used the *c-rater*™ (Leacock & Chodorow, 2003) automated scoring engine to analyze and score responses to these questions. Developed by ETS, c-rater is a scoring engine designed to evaluate short-answer CRs to questions that measure understanding of instructional content. C-rater differs from essay scoring systems like *e-rater*® (Attali & Burstein, 2006) in several fundamental ways. One is that c-rater's primary task is to recognize paraphrase or equivalent meaning, whereas e-rater's main goal is to analyze the quality of the structure and form of the essay. Another difference is that c-rater scoring is based on a model of the correct answer that is created by content experts for each specific question. C-rater attempts to map student responses onto the experts' models in order to determine their correctness or adequacy. E-rater scoring, on the other hand, can be based on a generic model of writing that is applied to any prompt that belongs to an assessment.

An observer looks directly at an object under a low level of illumination and the object is not seen. However, when the observer's eyes are shifted slightly so that the object is in peripheral vision, the object becomes visible because

(A) the rods are more sensitive to light than are the cones

(B) the cones are more sensitive to light than are the rods

(C) the cones enable greater acuity than do the rods

(D) visual acuity is better in foveal vision

(E) visual acuity is better in peripheral vision

***Figure 1.* Example question from the *GRE Psychology Test Practice Book* (p. 8), 2001, Princeton, NJ: ETS. Copyright 2001 by ETS. Used with permission.**

A c-rater item model can specify several related concepts that the student response should include in order to be viewed as complete. For example, the expert model could include the following information about photoreceptors: there are two kinds of photoreceptors; they are located on the retina; their names are cones and rods; cones are primarily located on the center part of the retina; rods are located on the peripheral part of the retina; rods are more sensitive to light than cones; cones are more sensitive to color than rods; in poor light conditions visual acuity is better in peripheral vision.

C-rater will identify the presence of each concept that is included in the student response. The item model can also include concepts that should not appear in the response. Scoring of the response can be guided by a complex set of rules about the appearance or absence of any concept in the model.

C-rater has been used in two studies to score CR reading comprehension items (Leacock & Chodorow, 2003). In both a National Assessment of Educational Progress (NAEP) study and a statewide assessment in Indiana, c-rater agreed with human raters about 84% of the time. A related study for automatically scoring analytical reasoning tasks was done by Kaplan and Bennett (1994). The study explored the potential of using an automated tool for the formulating-hypotheses item type. In this item, the student is presented with a situation and asked to generate explanations for it. In this study, correlations between program and human rater item scores ranged from .89 to .97.

In these studies, automated scoring was performed in batch mode following the administration. In this study, c-rater was applied for the first time immediately after participants submitted their responses. With on-the-fly scoring, it was possible to provide immediate feedback on the correctness of responses according to their coverage of the model answer. For responses that covered only some of the key concepts in the model answer, appropriate feedback was provided that indicated to students their answer was not complete (see the appendixes for example questions and models). When a response was not awarded full credit (either partial or no credit), a second attempt was allowed and participants were encouraged to correct their answer.

A fundamental question in this study concerns the quality of automated scoring with c-rater. To answer this question, all student responses were subsequently scored by two human raters and the agreement between human raters as well as between human and machine ratings was analyzed. This question is the focus of this report.

A separate report (Attali & Powers, 2008) focuses on the feedback and revision manipulation and addresses whether participants were able to correct their initial answers following c-rater’s feedback, whether the reliability of the revised scores were higher than the reliability of the scores based on initial answers, and whether the correlations of the revised scores with the GRE Subject Test scores were higher than the correlation of the initial scores.

Method

Participants

Study participants were recruited from GRE test registration files. Students who registered for either the Biology or Psychology Subject Test in November or December of 2005 were sent an invitation letter to participate in a Web-based study to evaluate experimental item types. The letters were sent 3 weeks before the test dates (November or December). In addition, those candidates for whom an e-mail address was present in the registration files received a second e-mail invitation 5 days prior to the test date. As an incentive, five $100 gift certificates were promised to be randomly distributed among study participants. Overall, the number of invitations sent was around 10,670, and around 3,400 of them were Biology Test applicants (32%). A total of 971 participants completed the study (9% of total invited students), 331 biology and 640 psychology students.

Out of the 971 participants, 919 (95%) had valid GRE Subject Test scores from the November-December administration. A comparison of the study participants with the general population of Subject Test test-takers (ETS, 2006) shows that the percent of women in the study (66% and 81% for biology and psychology, respectively) is similar to that in the general population of Subject Test test-takers (65% and 77%). The GRE Biology Subject Test scores of the study participants (*M* = 698, *SD* = 111) are high compared to the general population (*M* = 647, *SD* = 117). The GRE Psychology Subject Test scores of the study participants (*M* = 631, *SD* = 89) are also high compared to the general population (*M* = 592, *SD* = 101).

Materials

A total of 15 open-ended biology questions were adapted, by test developers for the GRE Biology Subject Test, from MC items included in an older edition of the official GRE guide *Practicing to Take the Biology Test* (3rd edition from 1994). Similarly, 20 open-ended psychology questions were also adapted. The questions were selected on the basis of an intuitive evaluation of their suitability for c-rater analysis and a desire to cover a wide range of subtopics from the original test.

In order to collect possible answers for c-rater model building, a pilot was conducted in April 2005. Registered students for the Biology and Psychology Subject Tests in April 2005 were invited, via e-mail, to answer the questions by replying to the invitation e-mail. The incentive for answering and replying was $100 gift certificates awarded to one out of each 50 participants. In addition, participants were e-mailed back with suggested correct answers to the questions. Overall, 58 students responded with answers to the biology questions and 101 with answers to the psychology questions. These represent a small sample size for c-rater model building, especially for the biology questions. The answers to the questions were reviewed by the test developers to determine if c-rater model building was viable. For psychology, 14 out of the original 20 questions were judged as possible candidates for model building. For biology, 12 out of the original 15 questions were judged as possible candidates for model building.

C-rater model building proceeded by first developing, for each question, a scoring and feedback model that specified feedback categories (and accompanying feedback text) that would be presented for each response, and how these feedback categories would be mapped into scoring categories of full, partial, and no credit (see the appendixes for examples). For all but four psychology questions a partial credit scoring category was included. Since the number of real responses available was small the c-rater feedback models were not based solely on existing responses. Test developers also tried to anticipate future responses based on their expertise in the field.

After these feedback and scoring models were completed the available responses were scored according to these models and were used in the development of the c-rater scoring models. Again, because of the small number of available responses, some of the c-rater feedback categories were not presented in the existing set of responses. The c-rater models were developed on the basis of all available responses, and no validation set of responses was used in the development process. This feature is also unusual in c-rater procedures.

In the process of c-rater model building it became clear that limitations of the c-rater scoring engine would not enable c-rater scoring for one biology question and three psychology questions. Consequently the two final test forms included 11 biology and 11 psychology questions.

Procedure

These items were administered to registered GRE applicants 1 to 2 weeks before their operational test. The study used a Web-based delivery system that allowed participants to take the test from any Internet-connected computer. The students used their Web browser to navigate to a login page that was included in the invitation letter. After reading an initial introduction and general instructions page, the students answered the 11 questions in their form in a fixed order.

When a student submitted his or her answer to a question, the c-rater engine analyzed the response and assigned it to a feedback category. The corresponding feedback text was then presented to the student. If the response was categorized as incomplete (less than full credit) the student was asked to revise the answer and submit it again. This revised response was analyzed in the same way and feedback was provided again, but without further opportunities to revise the answer. This process was repeated for each of the 11 questions. Following the test, the participants were asked several questions about their attitude towards these CR items and about their level of confidence in guessing answers to MC questions. Finally, a report with the student’s final answers and model answers for each question was presented to the student.

After test administration ended, all participant responses were submitted to human expert evaluation. The two pairs of test developers (one pair for each subject area) who were involved in the c-rater model development independently assigned feedback categories to each response. These feedback (and credit score) assignments were used to compare human and machine scoring.

Results

The results section starts with an analysis of human and machine agreement at the level of feedback categories. This level of analysis is more detailed than the score level, since several feedback categories could be assigned the same score, but it is important in a diagnostic application of c-rater. Next, the human and machine agreement at the level of credit scores is analyzed, followed by an analysis at the test level (reliability and GRE correlations). This section concludes with an analysis of participant attitudes and response length.

Analysis of Feedback Categories

The Biology scoring models provided more detailed feedback than the Psychology models. Table 1 shows that, for Psychology models, the average number of different feedback categories was just over one per credit level. For Biology models it was 2 for partial credit and almost 3 for no credit.

Tables 2–3 present agreement results for human and automated scoring (first attempt only) with regard to feedback categories. Exact agreement between H1 and H2 is presented as reference and the kappa between all three pairs of raters are compared.

Table 1

Number of Feedback Categories Across 11 Questions

Credit	Biology	Psychology
None	29	11
Partial	22	13
Full	11	14
Overall	62	38

Table 2

Biology Feedback Assignment Agreement, First Attempt

Question	Exact agr. H1-H2	Kappa H1-H2	Kappa H1-CR	Kappa H2-CR
1	.98	.97	.87	.90
2	.87	.81	.65	.68
3	.86	.78	.45	.44
4	.82	.76	.55	.60
5	.80	.73	.34	.39
6	.73	.64	.47	.51
7	.83	.79	.68	.77
8	.63	.53	.47	.68
9	.94	.80	.51	.59
10	.89	.78	.49	.54
11	.86	.79	.52	.52
Average	.84	.76	.54	.60

Note. N = 331. H = human-scored, CR= c-rater-scored.

The tables show that the human-human agreement is lower for biology questions (an average kappa of .76 compared to .89 for psychology questions), maybe due to the higher number of feedback categories for biology questions. However, following Landis and Koch’s (1977) suggestions about the interpretation of kappa values, both averages represent excellent agreement beyond chance (their values are greater than .75). The tables also show that the advantage of human-human agreement over human-machine agreement is larger for biology questions (about 19 kappa points on average for biology compared to 9 points for psychology). However, most psychology human-machine kappa values represent excellent agreement, and most biology human-machine kappa values represent fair to good agreement (between .40 and .75). Only the human-machine kappa value of biology Question 5 can be characterized as poor agreement (below .40).

Table 3

Psychology Feedback Assignment Agreement, First Attempt

Question	Exact agr. H1-H2	Kappa H1-H2	Kappa H1-CR	Kappa H2-CR
1	.89	.78	.63	.67
2	.95	.93	.87	.91
3	.86	.79	.56	.58
4	.98	.96	.97	.97
5	.99	.98	.97	.99
6	1.00	.98	.95	.96
7	.88	.81	.75	.72
8	1.00	.99	.81	.82
9	.97	.92	.87	.88
10	.87	.84	.72	.74
11	.84	.76	.76	.73
Average	.93	.89	.81	.81

Note. N = 640. H = human-scored, CR= c-rater-scored.

Analysis of Credit Scores

In addition to the feedback that was provided for every response, all responses were awarded full, partial, or no credit. In the following analyses of credit scores, full credit was interpreted as a full point, partial credit as half a point, and no credit as 0 points. Tables 4–5 present descriptive statistics about the first attempt credit scores. The first three columns show the mean and, in parentheses, *SD* of the human scores (H1 and H2) and the automated c-rater scores (CR). The last columns show the weighted kappa (κ_w) results between the three pairs of scores. The weighting in the computation of kappa was linear, as suggested by Cicchetti and Allison (1971).

Results from the mean and *SD* columns of Tables 4-5 suggest that the automated scores are slightly lower than the human scores, by .04-.07 overall for biology and by .01-.02 overall for psychology. The effect sizes (*d*) for these differences are mostly small (.2-.5 of the standard deviation of scores, see Cohen, 1988) or insignificant (smaller than .2 of the standard deviation of scores).

Table 4

Mean (SD) and Kappa Measures for Biology Credit Scores, First Attempt

							κ_w	κ_w	κ_w
Question	H1		H2		CR		H1-H2	H1-CR	H2-CR
1	.81	(.32)	.81	(.32)	.81	(.33)	.97	.91	.92
2	.33	(.34)	.32	(.34)	.28	(.35)	.85	.66	.68
3	.41[c]	(.44)	.46	(.43)	.32[ab]	(.43)	.83	.48	.47
4	.46[c]	(.39)	.50	(.40)	.39[ab]	(.42)	.82	.60	.62
5	.37	(.39)	.39	(.40)	.32[b]	(.38)	.82	.45	.50
6	.47[c]	(.45)	.55	(.40)	.49[b]	(.40)	.70	.52	.57
7	.43	(.37)	.46	(.34)	.44	(.34)	.82	.73	.80
8	.56	(.40)	.57	(.38)	.55	(.40)	.69	.61	.77
9	.85	(.32)	.85	(.33)	.75[ab]	(.43)	.85	.57	.64
10	.72	(.41)	.75	(.41)	.67[b]	(.43)	.82	.59	.61
11	.37	(.32)	.40	(.33)	.27[ab]	(.32)	.80	.52	.52
Overall	.52[c]	(.42)	.55	(.41)	.48[ab]	(.43)	.84	.66	.69

Note. $N = 331$. H = human-scored, CR= c-rater-scored.

[a] $p < .001$ for the comparison between H1 and CR. [b] $p < .001$ for the comparison between H2 and CR. [c] $p < .001$ for the comparison between H1 and H2.

The kappa values for credit scores are in general higher than those for feedback categories and the pattern of results between raters and subject area is similar for feedback and credit scores.

Score Revision

An important question in the context of this study is to what extent participants can correct their previous incorrect answers in response to the feedback they received. This question is explored more fully in the companion report (Attali & Powers, 2008), whereas in this report we focus on the relation between human and machine scores for revised responses.

Table 5

Mean (SD) and Kappa Measures for Psychology Credit Scores, First Attempt

							κ_w	κ_w	κ_w
Question	H1		H2		CR		H1-H2	H1-CR	H2-CR
1	.48	(.50)	.52	(.50)	.48	(.50)	.78	.63	.67
2	.27	(.31)	.28	(.32)	.30[ab]	(.33)	.94	.88	.91
3	.42	(.44)	.45	(.44)	.49[ab]	(.46)	.85	.76	.79
4	.35	(.48)	.36	(.48)	.35	(.48)	.96	.97	.97
5	.65	(.43)	.65	(.43)	.65	(.43)	.99	.98	.99
6	.88	(.32)	.88	(.33)	.87	(.33)	.98	.95	.96
7	.34[c]	(.42)	.39	(.43)	.33[b]	(.41)	.84	.80	.78
8	.14	(.31)	.14	(.31)	.10[ab]	(.26)	.99	.75	.77
9	.22	(.41)	.22	(.42)	.19[ab]	(.39)	.92	.87	.88
10	.55	(.34)	.56	(.35)	.53	(.33)	.91	.74	.76
11	.60	(.32)	.60	(.35)	.58	(.32)	.77	.77	.74
Overall	.45[c]	(.45)	.46	(.45)	.44[ab]	(.44)	.92	.86	.87

Note. N = 640. H = human-scored, CR= c-rater-scored.

[a] $p < .001$ for the comparison between H1 and CR. [b] $p < .001$ for the comparison between H2 and CR. [c] $p < .001$ for the comparison between H1 and H2.

Therefore, Table 6 shows only a summary of the success of participants in revising their initial, less-than-full-credit scores. It shows that out of the cases where participants did not receive full credit in their first attempt, around 80% took advantage of the second attempt to revise their answer. In the cases where no credit was given in the first attempt, 27% (in biology) and 16% (in psychology) improved their scores. In the cases where partial credit was given in the first attempt, around 23% improved their scores but for around 10% the revised score was lowered.

Although Tables 2-5 showed that the agreement between automated and human scores was relatively high for first attempt responses, it is interesting to compare the human and machine scores of revised answers. Tables 7-8 present this comparison. Similarly to the first attempt results (shown in Tables 4-5) the automated scores are slightly lower than the human scores. We can conclude that the pattern of results for human and machine scores in the first and second attempts is similar.

Table 6

Frequencies of Revised Automated Credit Scores

		1st Attempt – No credit			1st Attempt – Partial credit		
Test	Revised 2nd attempt	No	Partial	Full	No	Partial	Full
Biology	82%	73%	15%	12%	8%	69%	22%
Psychology	76%	84%	6%	11%	11%	65%	24%

Another analysis that sheds light on the number of errors in feedback and their consequences with revised responses is presented next. This analysis shows the percentage of cases where the two human raters gave full credit to the first answer of a respondent, whereas c-rater did not give full credit to this answer. Consequently, the respondent was asked to revise his or her answer. The question is how many of these (presumably) erroneous cases resulted in a revised answer that was judged by the human raters to deserve less than full credit, thus weakening a previously correct answer.

Table 7

Biology—Mean (and SD) Credit Scores, Second Attempt Offered

Question	*N*	H1	H2	CR
1	91	.57 (.41)	.56 (.41)	.51 (.44)
2	291	.34 (.38)	.35 (.38)	.30 (.36)
3	249	.29 (.40)	.38 (.43)	.25 (.38)
4	244	.45 (.39)	.48 (.40)	.36 (.39)
5	275	.30 (.36)	.31 (.36)	.26 (.30)
6	228	.38 (.41)	.47 (.39)	.36 (.36)
7	275	.43 (.35)	.45 (.31)	.42 (.27)
8	208	.50 (.43)	.51 (.38)	.46 (.35)
9	85	.62 (.45)	.61 (.46)	.40 (.49)
10	131	.53 (.45)	.56 (.47)	.41 (.44)
11	305	.37 (.32)	.40 (.33)	.33 (.33)
Overall	2,382	.40 (.39)	.43 (.39)	.35 (.37)

Note. H = human-scored, CR= c-rater-scored.

Table 8

Psychology—Mean (and SD) Credit Scores, Second Attempt Offered

Question	*N*	H1	H2	CR
1	335	.33 (.47)	.36 (.48)	.25 (.43)
2	580	.26 (.29)	.26 (.29)	.29 (.29)
3	375	.23 (.38)	.27 (.40)	.28 (.39)
4	413	.15 (.36)	.15 (.36)	.14 (.35)
5	279	.27 (.39)	.27 (.39)	.26 (.38)
6	81	.59 (.49)	.62 (.49)	.60 (.49)
7	497	.25 (.36)	.31 (.39)	.23 (.34)
8	607	.10 (.28)	.10 (.28)	.07 (.23)
9	520	.13 (.34)	.13 (.34)	.09 (.29)
10	481	.56 (.40)	.57 (.40)	.51 (.37)
11	450	.54 (.29)	.54 (.31)	.52 (.26)
Overall	4,618	.28 (.39)	.29 (.40)	.26 (.37)

Note. H = human-scored, CR= c-rater-scored.

For biology, out of 3,641 first answers, 207 (5.7%) were awarded full credit by both human raters but less than full credit by c-rater. Out of these 207 cases, 50 (24%) of the second answers were awarded less than full credit by at least one of the human raters. Overall, these 50 cases constitute 1.4% of all first answers. For psychology, out of 7,040 first answers, 230 (3.3%) were awarded full credit by both human raters but less than full credit by c-rater. Out of these 230 cases, 50 (22%) of the second answers were awarded less than full credit by at least one of the human raters. Overall, these 50 cases constitute 0.7% of all first answers. Overall, about 1% of all first answers resulted in unnecessarily asking the respondent to correct their answer, and the correction seemed to harm the initial answer to some degree.

Score Reliabilities and GRE Correlations

This section compares the study score reliabilities and correlations with GRE Subject Test scores between human and machine scores. The automated scores that are analyzed are based on the first answer, last answer, and highest score across the one or two attempts. Obviously, if the first answer received full credit the first, last, and highest scores are the same for this question. Otherwise, the last-answer score will be equal to the score in the second attempt, and the highest-answer score will equal the highest of the two.

Table 9 shows that Cronbach's coefficient alpha estimates of reliability for the first answer automated scores are lower than the human score reliabilities for both tests, but not significantly different than the human reliabilities. However, last-answer (for biology) and highest-answer (for both tests) score reliabilities are significantly higher than human score reliabilities. The comparison of human and machine correlations with GRE scores shows that in all cases the machine correlations are lower, but only the first-answer biology correlation is significantly different than the human correlation.

Table 9

Study Score Reliabilities and Correlations with GRE Subject Test Scores

	α	r
Biology		
Human score[a]	.568	.57
First answer	.532	.50*
Last answer	.639*	.53
Highest answer	.649*	.52
Psychology		
Human score	.607	.59
First answer	.590	.58
Last answer	.637	.56
Highest answer	.650*	.56

*Note. N*s are 331 and 640 for reliabilities, 320 and 599 for GRE correlations.

[a] Average results for H1 and H2.

* $p < .05$ for the dependent t-test comparison with human value.

Examinee Attitudes

The first six questions in the post-test questionnaire referred to student attitudes towards the different test formats. The first survey question was "Was it helpful to know, before you answered a question, that you were going to get feedback on your answer?” The response options were definitely yes, probably yes, not sure, probably no, and definitely no. The relative distribution of responses was 49%, 38%, 5%, 6%, and 2%, respectively, showing that respondents clearly felt knowing they were getting feedback was helpful. Most participants also thought the feedback helped in correcting initial answers “to some degree” (66%) or “considerably” (11%), whereas only 23% thought “it did not help at all.”

Nevertheless, when asked how many times they felt the feedback was incorrect only 15% answered feedback was always correct. The percentage of respondents selecting 1 to 6 times as the number of incorrect feedback occurrences was 24%, 23%, 15%, 10%, 7%, and 2%. A total of 5% of respondents chose 7 to 11 times.

A large majority (82%) preferred a MC test to an open-ended test, 16% preferred an open-ended test with feedback and revision, and 1% preferred an open-ended test without feedback. This may be explained by the responses to the question about which kind of test is least stressful, where again a large majority (81%) chose the MC test (13% chose the open-ended test with feedback and 6% chose the open-ended test without feedback). However, a majority of participants think that a better indicator of their ability is an open-ended test, either with feedback and revision (51%) or even without feedback (14%). Bridgeman (1991) similarly found that examinees preferred the MC format to an open-ended format (without feedback) but were split about the perceived fairness of these two formats.

Length of Responses

The purpose of this section is to analyze the relation between response length, measured here as number of words, and response score. Response length is a very important predictor of scores in the context of essay writing, and it is interesting to compare its importance to the context of short questions that are content based.

Response lengths for the first attempts were, overall, slightly lower in psychology (M = 14.2, SD = 15.8) than in biology (M = 17.2, SD = 15.2). The overall correlation between credit scores (first attempt only), both human and machine, and response length was computed for each subject area. For biology the correlations were .21, .19, and .22 for c-rater, the first human rater, and the second human rater. For psychology the correlations were .13, .10, and .12 for c-rater, the first human rater, and the second human rater. These results suggest that there is an overall positive relation between response length and scores, although the relation is not particularly strong. They also show that the c-rater scores are not more highly correlated with response length than human scores.

The purpose of the second analysis was to see whether the average response length for an item is related to the score-length correlation that was observed in a general way. Figure 2 is a scatter plot, for the 22 items, of the item-level score-length correlation (with a median of .24) against the item-level average response length (with a median of 15 words). The figure shows that higher score-length correlations are associated with items that elicited longer responses.

Summary

The feasibility of employing the c-rater engine for scoring responses on-the-fly was evaluated in this study. The types of questions that were used in this study, open-ended variants of Subject Test MC questions in biology and psychology, are a good fit for the c-rater engine. They have a strong content basis, focus on a limited set of concepts, and require a short answer of 1-3 sentences. On the other hand, the limited number of training responses for the development of c-rater models (around 60 and 100 for biology and psychology, respectively) was a concern for the quality of models. Nevertheless, c-rater scoring was, overall, successful in both subject areas.

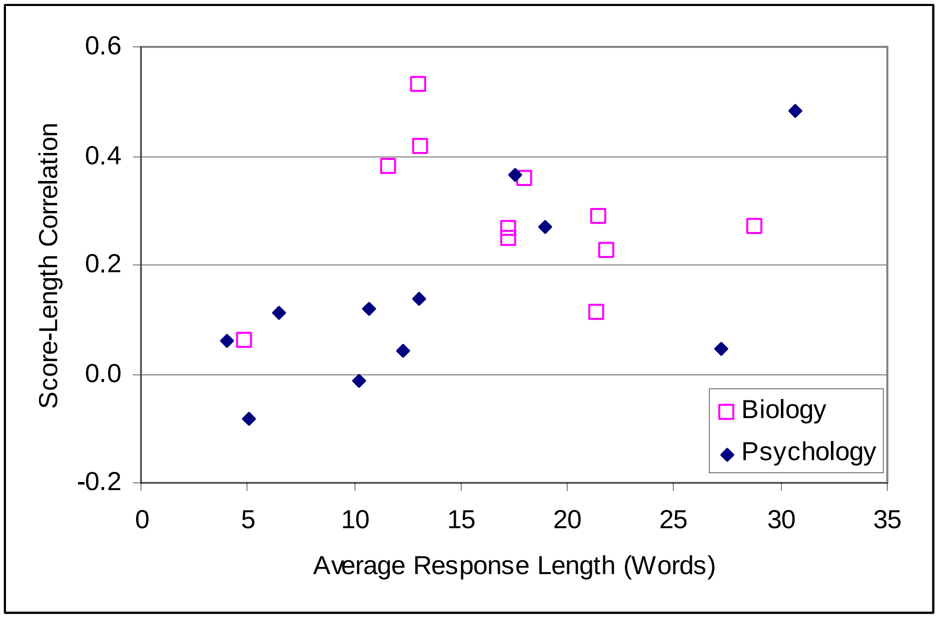

Figure 2. **Relation between average response length and score-length correlation.**

Both for assignment of feedback categories and for credit scores (full, partial, or no credit), the kappa agreement values between human and machine assignments were lower than the agreement between two human raters. The differences were more pronounced for biology questions (around .15-.20 points) than for psychology questions (around .10), but in general agreement was higher for psychology questions. Despite these differences, the agreement between human and machine assignments across questions represented at least fair to good agreement.

Another indication that machine scoring did not deteriorate the quality of scores in a significant way comes from the test reliability and GRE correlation analysis. For both tests, the test reliability of first attempt scores based on machine scoring was not significantly different than reliability based on human scoring. And only for the biology test, the GRE correlation of first attempt c-rater scores was significantly lower than the correlation with human scores. Moreover, when taking into account revised scores, the c-rater test reliabilities were significantly higher than the human reliabilities. This is significant because only through automated scoring is it possible to provide immediate feedback and elicit revised answers that would increase the reliability of scores in this way.

Although most respondents felt that automated feedback was not always accurate, most also felt that, overall, it was helpful to know that feedback will be provided, and that feedback helped correct mistakes. As a matter of fact, most respondents (around 80%) also chose to revise their answers following feedback that suggested so. This is another indication that, overall, automated scoring was successful.

Finally, another encouraging result was the rather low correlations (with a median of .24 across all questions) between response length and both human and machine scores. The analysis also confirmed that questions that are associated with longer responses also tend to produce higher correlations between response length and scores.

In conclusion, although the c-rater engine is not suitable to score just any question and its use requires a considerable effort in modeling responses, the use of c-rater was generally successful in this study.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved March 21, 2008, from http://escholarship.bc.edu/jtla/vol4/3/

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 GRE-04-05).Princeton, NJ: ETS.

Bridgeman, B. (1991). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*, 253-271.

Briel, J. B., O'Neill, K. A., & Scheuneman, J. D. (Eds.). (1993). *GRE technical manual: Test development, score interpretation, and research for the Graduate Record Examinations Program* (pp. 67-88). Princeton, NJ: ETS.

Cicchetti, D.V., & Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology, 11*, 101-109.

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

ETS. (2006). *GRE guide to the use of scores*. Princeton, NJ: Author.

Glanz, J. (1996). How not to pick a physicist? *Science, 274*, 710-712.

GRE psychology test practice book. (2001). Retrieved March 21, 2008, from http://ftp.ets.org/pub/gre/Psychology.pdf

Kaplan, R. M., & Bennett, R. E. (1994). *Using the free-response scoring tool to automatically score the formulating-hypotheses item* (ETS Research Rep. No. RR-94-08). Princeton, NJ: ETS.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2000). Comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*, 162-181.

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

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

Notes

[1] Marissa Harrison is currently at Borough of Manhattan Community College, The City University of New York.

Appendix A

Example Biology Question

Original MC Item

Colchicine, which has long been used by plant geneticists to produce artificial polyploids, is known to act by

(A) attaching to the centromeres at metaphase so that anaphase cannot occur

(B) binding to the protein tubuline and preventing the assembly of spindle microtubules

(C) inhibiting the microfilaments and preventing cleavage or furrowing

(D) preventing the replication of centrioles and spindle formation

(E) separating the strands of the DNA double helix and stimulating chromosome duplication

CR Version

Describe how colchirine acts to produce artificial polyploids in plants.

Model CR Answer

Colchicine prevents microtubules from forming correctly. Since microtubules play important roles in chromosome separation during meiosis, if colchicine is introduced to plant cells undergoing meiosis, chromosomes might not separate and polyploids, or offsprings with too many sets of chromosomes, could result.

Full Credit Feedback

You have provided a correct answer.

No Credit Feedback

You have not provided a correct or sufficient response.

Partial Credit Feedback

If response only mentions the prevention of chromosome separation and/or stops in mitosis or meiosis, feedback reads:

You have described some of the effects of colchicine addition, but you have not described how it specifically acts to produce artificial polyploids.

If response only mentions that colchicine inhibits microtubules/spindle, feedback reads:

You have described an action of colchicine, but you have not described how this action produces artificial polyploids.

Appendix B

Example Psychology Question

Original MC Item

The way in which certain birds (such as white-crowned sparrows) learn their song and the way in which ducklings learn to follow their mother are similar in which of the following respects?

(A) Both occur readily in the natural environment but cannot be demonstrated under laboratory conditions.

(B) Both require positive reinforcements, which the birds' parents provide.

(C) Both produce short-lived effects that influence behavior for only a week or two.

(D) Both are learned most easily during a sensitive period of development.

(E) Both are facilitated by androgens and inhibited by estrogens.

CR Version

The way in which certain birds (such as white-crowned sparrows) learn their song and the way in which ducklings learn to follow their mother are similar in what respects?

Model CR Answer

Birds learn their songs and newborn ducklings learn to follow their mothers through imprinting, which is rapid, relatively permanent learning that occurs during a critical period/sensitive period regardless of the consequences of a behavior.

Full Credit Feedback

You have provided a correct answer.

No Credit Feedback

You have not provided a correct or sufficient response.

Partial Credit Feedback

If response only mentions the concept of imprinting, feedback reads:

You correctly identified imprinting as a common principle that the examples share. What is another essential psychological principle that they share?

If response only mentions the concept of critical period, feedback reads:

You correctly identified a critical/sensitive period as a common principle that the examples share. What is another essential psychological principle that they share?

GRE-ETS
PO Box 6000
Princeton, NJ 08541-6000
USA

To obtain more information about GRE
programs and services, use one of the following:
Phone: 1-866-473-4373
(U.S., U.S. Territories*, and Canada)
1-609-771-7670
(all other locations)
Web site: www.gre.org

* America Samoa, Guam, Puerto Rico, and US Virgin Islands