Effect of Immediate Feedback and Revision on Psychometric Properties of Open-Ended GRE Subject Test Items

Yigal Attali
Don Powers

April 2008

ETS GRE Board Research Report No. 04-05
ETS RR-08-21
Effect of Immediate Feedback and Revision on Psychometric Properties of Open-Ended GRE® Subject Test Items

Yigal Attali and Don Powers
ETS, Princeton, NJ

GRE Board Research Report No. 04-05
ETS RR-08-21

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.

ETS, the ETS logos, GRADUATE RECORD EXAMINATIONS, GRE, and LISTENING. LEARNING. LEADING. and are registered trademarks of Educational Testing Service (ETS). C-RATER is a trademark of ETS.

Educational Testing Service
Princeton, NJ 08541

Copyright © 2008 by ETS. All rights reserved.
Abstract

Registered examinees for the GRE® Subject Tests in Biology and Psychology participated in a Web-based experiment where they answered open-ended questions that required a short answer of 1-3 sentences. Responses were automatically scored by natural language processing methods (the c-rater™ scoring engine) immediately after participants submitted their responses. Based on natural language processing methods (the c-rater scoring engine), participants received immediate feedback on the correctness of their answers and an opportunity to revise their answers. A significant revision effect was found. Participants were able to correct many of their initial incorrect answers, resulting in higher revised scores. In addition, the reliability of revised scores was higher than initial scores, although the correlations of the initial and revised scores with the GRE Subject Test scores were similar.

Key words: Feedback, constructed response, GRE Subject Test
Introduction

Feedback in Assessment

Providing feedback regarding task performance is one of the most frequently applied of all psychological interventions (see Kulhavy & Stock, 1989, for a review). Thorndike (1913) provided the initial theoretical arguments for the effectiveness of feedback with his law of effect, in which feedback was regarded as a contingent event that reinforced or weakened responses. More recently, the cognitive revolution has provided a new framework for understanding feedback’s role in learning, highlighting its informational role. From this perspective, errors are viewed as a source of information about students’ cognitive processes (see Bruning, Schraw, & Ronning, 1999). Feedback helps learners determine performance expectations, judge their level of understanding, and become aware of misconceptions. It may also provide clues about the best approaches for correcting mistakes and improving performance.

Considering its major role in learning and instruction, it is curious that feedback has had almost no place in assessment. In this respect, it is helpful to distinguish between knowledge of one’s performance and allowing revision of one’s response. Although this distinction may not be entirely relevant for learning and instruction, it is commonplace in testing environments to provide one or the other (or none) but not both.

The option to review and change answers is taken for granted on most paper-and-pencil tests, but for technical reasons it is usually impractical to provide item-by-item feedback on performance. Researchers have studied answer review and change behavior on paper-and-pencil tests for some time (for a review, see Waddel & Blankenship, 1995). As a whole, this research shows that a small percentage of answers are typically changed, more answers are changed from wrong to right than from right to wrong, and a large proportion of examinees change answers to at least some questions.

Surprisingly, even answer review by itself is less common on computerized tests, especially when the test items are administered adaptively based on the examinee’s performance on previously answered items. Arguments against the use of item review on computerized adaptive tests (CATs) have focused mostly on increases in testing time, possible reductions in measurement precision, complications in item administration algorithms, and reduced test score validity resulting from answer response strategies that yield inflated ability estimates (Wainer, 1993).
However, in general examinees strongly favor answer review, presumably because it provides a less stressful testing experience (Vispoel, 2000).

In the context of computerized tests, knowledge of results, but without answer review, has also been studied (Betz, 1977). One reason that the combination of immediate knowledge of results and answer review is not provided in computerized (non-adaptive) tests may be that most of these tests are in multiple-choice (MC) format. With this item format, providing feedback on the correctness of an answer raises the probability of answering correctly in a subsequent trial simply as a result of the decreasing number of options.

An exception is the answer-until-correct scoring method (Wilcox, 1982) that provides both immediate feedback and allows answer review for MC items by providing partial credit based on the number of trials until the correct option for a MC item is selected. However, this scoring method is rarely used in assessments. It is also interesting to note that the MC item format itself inherently provides negative feedback, because if the examinee does not find her answer in the set of options, she knows her answer is not correct.

In summary, traditional assessments rarely provide immediate feedback about the correctness of an answer with an opportunity to revise answers. However, even in the context of the tutoring literature where feedback is a major tool for enhancing learning, feedback is typically provided in an acquisition or training phase but not when the effects of learning are assessed. For example, Schmidt, Young, Swinnen, and Shapiro (1989) investigated the timing of feedback with respect to the acquisition of motor skills. Subjects received feedback about their performance following a variable number of trials in a motor learning task, but when learning was assessed no feedback was provided. Schooler and Anderson (1990) investigated the same timing of feedback question in the context of learning computer programming skills. During training subjects received different kinds of feedback but when the effects of these kinds of feedback were tested no feedback was provided.

Finally, dynamic assessment (for a review, see Grigorenko & Sternberg, 1998) provides a unique view of feedback. In contrast to standard assessments, which are seen as static, dynamic assessments are designed to explore the potential of the student for learning. Such assessments are based on an interactive approach to assessment that embeds intervention within the assessment procedure. This approach is related to Vygotsky's (1934/1986) concept of the Zone of Proximal Development, the gap between what can be learned unaidered and what can be learned with the help
of a more capable peer. A common procedure used in dynamic assessment is the test-teach-retest procedure, whereby the difference in performance between the two tests is the focus of assessment. For example, Embretson (1992) found support for the validity of cognitive modifiability for predicting learning in a spatial domain. However, as in the tutoring literature, the tests used as the basis for assessing modifiability are static (MC in Embretson, 1992) and do not provide immediate feedback and opportunity to revise answers.

In the tradition of dynamic assessment, the graduated-prompts approach of Campione and Brown (1987) present a different approach to within-test feedback. Whereas other practitioners of dynamic assessment (e.g., Feuerstein, 1979) were primarily concerned with the amount of improvement in performance following guided instruction, the graduated-prompts approach focused on measuring the amount of instruction needed to bring students to some specified level of competence. This was operationalized through a guided problem-solving paradigm, where students receive a gradually increasing series of hints until they are able to solve the problem. The hints consist of both task-specific and general meta-cognitive guidance, and are quite detailed. The measure of learning potential is defined as the inverse of the minimal number of hints that the student needs to reach a specified performance level (e.g., solving three consecutive problems unassisted). Studies in this paradigm (e.g., Day, Engelhardt, Maxwell, & Bolig, 1997; Ferrara, Brown, & Campione, 1986) showed that learning potential measures predicted performance gains from training beyond the contribution of general ability or initial status.

**An Approach for Interactive Assessment Based on Correctness Feedback**

The advance of technology and computerized testing facilitates the provision of immediate feedback to examinees by making it more feasible to use constructed-response (CR) tasks that can be scored automatically on-the-fly. In such a CR testing environment it is possible to inform students that an initial answer is incorrect and ask them to revise it, without encountering the problems noted above for MC tests. The purpose of this study, as well as a previous study by Attali, Powers, and Hawthorn (2007), was to investigate the psychometric effects of providing both immediate feedback on the correctness of an answer and opportunity to revise incorrect answers.

The major difference between this approach and the graduated-prompts dynamic assessment of Campione and Brown (1987) has to do with the nature of feedback that is provided to the student. In our approach, only limited information on the correctness of the student response
is given, and no attempt is made to offer increasingly helpful hints. In addition, only a limited and fixed number of opportunities are given to revise an answer—one in this study and two in the Attali et al. (2008) study. The advantage of this kind of feedback is that it is easier to produce than are customized hints, and it provides standardized feedback across tasks and content areas. The standardized feedback and fixed number of revision opportunities also facilitate scoring comparability across items and tests.

Two general effects are of interest in this context. The feedback effect is the possible effect, on test-taking attitudes and behavior, of knowing that a second opportunity for answering the question will be given, prior to the first attempt. This effect can be evaluated by comparing the first attempt answers of examinees under feedback and no-feedback conditions. Such an effect would primarily be affective in nature. A test taker facing an open-ended question without an opportunity to correct an initial error could be more anxious than one who knows that revision will be allowed, because failure in the first attempt is less consequential when multiple attempts are allowed.

The second effect is the revision (or feedback and revision) effect, the effect of receiving feedback on a previous attempt and trying to correct an answer. This effect can be evaluated by comparing the answers of examinees (under feedback condition) in consecutive attempts. Such an effect would primarily be cognitive in nature. Although feedback on the correctness of a response does not provide any direct guidance for correcting the response, it could nonetheless bring about a reevaluation of the question and encourage students to explore alternatives for solving the problem.

Both the feedback and revision effects could have an impact on different aspects of test scores. The two primary questions are whether students are able to correct initial answers, and thus raise test scores, and whether test scores in a feedback and revision environment are more valid than test scores derived without feedback.

**The Vocabulary Study**

In a previous study, Attali et al. (2007) conducted an initial investigation of both the feedback and revision effects. Examinees registered for the GRE® General Test answered open-ended sentence-completion items. For half of the items participants received immediate feedback on the correctness of their answers and up to two opportunities to revise their answers. The other half of the items was administered without feedback and revision. With respect to the revision effect, the initial feedback scores were, surprisingly, significantly lower than no-feedback scores.
(although the standardized effect size, .14, was very small). Both the reliability and correlation with operational GRE Verbal scores were higher for initial feedback scores than for no-feedback scores, but the difference was significant only for the correlation.

With respect to the feedback effect, a significant and large effect was found for test scores. Scores based on revised answers were higher than initial scores (in the feedback condition) by .60 $SD$ after two attempts and by .93 $SD$ after three attempts. The reliabilities of scores based on revised answers were also significantly higher than initial scores. However, GRE correlations of initial and revised scores were not significantly different. In summary, participants were able to correct their initial answers and feedback, and revision had a beneficial effect on the quality of scores.

**The Present Study**

The purpose of this study was to replicate the feedback effect in different subject areas. Feedback and opportunity to revise answers were provided for all questions answered, and only one revision was allowed in this study. The questions answered by participants were content-based short-answer questions in biology and psychology that were developed as CR variants of MC questions in these two GRE Subject Tests. Examples of the original MC questions and the adapted CR versions are presented in Appendix A and B.

Scoring these CR 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 Technologies, c-rater is a scoring engine designed to evaluate answers to short CR questions that measure understanding of instructional content. C-rater’s primary task is to recognize paraphrase or equivalent meaning and 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.

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 for the example psychology question in Appendix B should include the concepts of imprinting and critical periods. 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 that will support partial scoring of incomplete responses as
well as specific diagnostic on-the-fly feedback to students according to their coverage of the model answer.

Naturally, 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. A complete analysis of c-rater’s success in scoring is reported elsewhere (Attali & Powers, 2008). The two research questions that are the focus of this study are whether participants would be able to correct their initial answers following c-rater’s feedback, whether the reliability of the revised scores would be 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 would be 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 sub-topics 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 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**

*Success in Automated Scoring*

All student responses in this study were awarded full, partial, or no credit by the c-rater engine, and a second attempt was provided in the case of partial or no credit. Following the test administration, all responses for each subject test were also scored by two content experts. A complete analysis of the success of c-rater in scoring these questions is provided in a separate
In this section, a summary of the agreement of c-rater scores with human scores is presented.

In the following analyses, full credit was interpreted as a full point, partial credit as half a point, and no credit as 0 points. Table 1 presents descriptive statistics about the first attempt credit scores and agreement results between the automated scores and the two human rater scores. The weighting in the computation of the weighted kappa measures was linear, as suggested by Cicchetti and Allison (1971).

Table 1 suggests 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 for these differences are insignificant (smaller than .2 of the standard deviation of scores). The weighted kappas between the two human raters are higher than between c-rater and the human raters, both for biology (by .15-.18) and psychology (by .05). Landis and Koch (1977) suggest that kappa values of .75 or higher should be interpreted for most purposes as excellent agreement, and values between .40 and .75 should be taken to represent fair to good agreement. According to this criterion, the overall level of agreement between automated and human scores is good for biology and excellent for psychology.

The weighted kappas are also higher for psychology than for biology (by .06 for human agreement and by almost .20 for human-machine agreement). This might be because the scoring models for biology questions tended to be more complicated.

<table>
<thead>
<tr>
<th>Test</th>
<th>N</th>
<th>H1</th>
<th>H2</th>
<th>CR</th>
<th>H1-H2</th>
<th>H1-CR</th>
<th>H2-CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td>331</td>
<td>.52c (.42)</td>
<td>.55 (.41)</td>
<td>.48ab (.43)</td>
<td>.84</td>
<td>.66</td>
<td>.69</td>
</tr>
<tr>
<td>Psychology</td>
<td>640</td>
<td>.45c (.45)</td>
<td>.46 (.45)</td>
<td>.44ab (.44)</td>
<td>.92</td>
<td>.86</td>
<td>.87</td>
</tr>
</tbody>
</table>

Note. H1, H2 = First and second human raters. CR = c-rater.
Success in Revision

The first major research 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. Tables 2–3 show the relative frequencies of revised (second attempt) automated credit scores for each first attempt automated score. The first column shows the percentage of students who did not receive full credit in their first attempt and thus were given a chance to revise their answer. The second column shows the percentage of students who actually changed their answer out of those who were given a second chance. If an answer was not revised it was awarded the same score in the second submission. The remaining columns show the percentages of students who received each credit score in the second attempt, for each first attempt score (no or partial credit).

Table 2
Biology—Percentages of Revised Automated Credit Scores

<table>
<thead>
<tr>
<th>Question</th>
<th>% 2nd chance</th>
<th>% Revised</th>
<th>1st Attempt—No credit</th>
<th>1st Attempt—Partial credit</th>
<th>1st Attempt—Full credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27</td>
<td>95</td>
<td>54</td>
<td>17</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>88</td>
<td>80</td>
<td>84</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>83</td>
<td>79</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>74</td>
<td>85</td>
<td>73</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>83</td>
<td>69</td>
<td>81</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>69</td>
<td>84</td>
<td>63</td>
<td>22</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>83</td>
<td>91</td>
<td>65</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>63</td>
<td>85</td>
<td>57</td>
<td>33</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>26</td>
<td>81</td>
<td>61</td>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>74</td>
<td>72</td>
<td>6</td>
<td>22</td>
</tr>
<tr>
<td>11</td>
<td>92</td>
<td>79</td>
<td>75</td>
<td>23</td>
<td>2</td>
</tr>
<tr>
<td>Overall</td>
<td>65</td>
<td>82</td>
<td>73</td>
<td>15</td>
<td>12</td>
</tr>
</tbody>
</table>

Note. N = 331.
Table 3

Psychology—Percentages of Revised Automated Credit Scores

<table>
<thead>
<tr>
<th>Question</th>
<th>% 2nd chance</th>
<th>% Revised</th>
<th>1st Attempt—No credit</th>
<th>1st Attempt—Partial credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1^a</td>
<td>52</td>
<td>87</td>
<td>75</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>91</td>
<td>81</td>
<td>84</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>59</td>
<td>66</td>
<td>85</td>
<td>9</td>
</tr>
<tr>
<td>4^a</td>
<td>65</td>
<td>79</td>
<td>86</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>44</td>
<td>82</td>
<td>70</td>
<td>17</td>
</tr>
<tr>
<td>6^a</td>
<td>13</td>
<td>83</td>
<td>40</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>78</td>
<td>78</td>
<td>84</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>95</td>
<td>66</td>
<td>94</td>
<td>3</td>
</tr>
<tr>
<td>9^a</td>
<td>81</td>
<td>68</td>
<td>91</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>75</td>
<td>80</td>
<td>84</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>70</td>
<td>82</td>
<td>57</td>
<td>35</td>
</tr>
<tr>
<td>Overall</td>
<td>66</td>
<td>76</td>
<td>84</td>
<td>6</td>
</tr>
</tbody>
</table>

Note. N = 640.

^a No partial credit for this question.

The tables show that the percentages of students who took advantage of the second attempt to change their answer were around 80% in most cases. The tables show considerable differences across questions, but in general, we can say that for those students who received no credit in their first attempt, around 10% of the revised answers were awarded partial credit (15% for biology and 6% for psychology) and 11% more were awarded full credit. For those students who received partial credit in their first attempt, around 10% (8% for biology and 11% for psychology) of the revised answers were awarded no credit (that is, their score went down) and around 23% (22% for biology and 24% for psychology) were awarded full credit. The two subjects seem to present similar results. An exception is that for psychology there is a lower percentage of partial credit for cases that were awarded no credit in the first attempt. However, this may be due to the fact that four psychology questions did not have partial credit.
**Test Scores**

Overall, the previous section showed that participants were able to correct their initial answers following automated feedback about the correctness of their answers. This section will evaluate the effects of revision at the level of test scores.

Table 4 presents descriptive statistics and reliabilities (Cronbach’s coefficient alpha) for different test scores. All analyses are based on actual automated scores computed during the test and not on “true” scores following the response review after administrations. The scores presented are based on the first answer, last answer, and highest score across all answers. 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 equal to the score in the second attempt, and the highest-answer score will equal the highest of the two.

Table 4 shows that the feedback and revision effect was significant and important. The last-answer scores were significantly higher than the first-answer scores in both subject areas, and the highest-answer scores were significantly higher than the last-answer scores in both subject areas. The effect sizes ($d$) for the last and highest scores were around .5 of the $SD$ of the first answer scores, a medium effect size.

**Table 4**

*Test Score Means (and SD) and Reliabilities (Based on 11 Items)*

<table>
<thead>
<tr>
<th></th>
<th>$N$</th>
<th>$M$</th>
<th>$SD$</th>
<th>$d$</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First answer</td>
<td>331</td>
<td>5.29</td>
<td>1.78</td>
<td>.532</td>
<td></td>
</tr>
<tr>
<td>Last answer</td>
<td>331</td>
<td>6.32**</td>
<td>2.00</td>
<td>.58</td>
<td>.639**</td>
</tr>
<tr>
<td>Highest answer</td>
<td>331</td>
<td>6.44**</td>
<td>1.97</td>
<td>.65</td>
<td>.649*</td>
</tr>
<tr>
<td>Psychology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First answer</td>
<td>640</td>
<td>4.86</td>
<td>1.91</td>
<td>.590</td>
<td></td>
</tr>
<tr>
<td>Last answer</td>
<td>640</td>
<td>5.68**</td>
<td>2.07</td>
<td>.43</td>
<td>.637**</td>
</tr>
<tr>
<td>Highest answer</td>
<td>640</td>
<td>5.80**</td>
<td>2.06</td>
<td>.49</td>
<td>.650**</td>
</tr>
</tbody>
</table>

* $p < .05$ for the dependent t-test comparison with previous value (above).

** $p < .01$ for the dependent t-test comparison with previous value (above).
Moreover, not only were the revised scores higher than the first-answer scores, they were also more reliable. Cronbach’s coefficient alpha estimates of reliability for the last-answer scores were significantly higher than the first-answer reliabilities in both subject areas, and the highest-answer reliabilities were significantly higher than the last-answer reliabilities in both subject areas.

The next analysis compares the correlations of the study scores, both human and automated, with GRE Subject Test scores. Table 5 shows these correlations together with the reliability estimates shown also in Table 4. The Ns are lower in this table because not all study participants were located in the GRE score files.

The first comparisons to observe in this table are for the GRE correlations between the different automated scores. For biology, the correlations of the revision scores (.53 and .52) are higher than the first-answer scores (.50) but the differences are not significant. For psychology, surprisingly, the correlations of the revision scores (.56) are lower than the first-answer scores (.58) but the differences are again not significant.

Table 5

<table>
<thead>
<tr>
<th>GRE Subject Test Correlations and Reliabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Biology</td>
</tr>
<tr>
<td>Human scoreb</td>
</tr>
<tr>
<td>First answer</td>
</tr>
<tr>
<td>Last answer</td>
</tr>
<tr>
<td>Highest answer</td>
</tr>
<tr>
<td>Psychology</td>
</tr>
<tr>
<td>Human score</td>
</tr>
<tr>
<td>First answer</td>
</tr>
<tr>
<td>Last answer</td>
</tr>
<tr>
<td>Highest answer</td>
</tr>
</tbody>
</table>

a Based on an estimate of .95 for the reliability of GRE scores.

b Average results for H1 and H2.
The comparison of GRE correlations between human and automated scores shows that the automated score correlations are lower, but the only significant difference is for the first-answer biology correlation (.50 versus .57, \( p = .01 \)).

Finally, an estimate of true-score correlations between the study scores and the GRE scores was computed by dividing the raw correlations by the square-root of the product of the score reliabilities. The estimate of the GRE Subject Test reliabilities was .95, based on the GRE Guide to the Use of Scores (Educational Testing Service, 2006). These estimates are quite low in general (.7 to .8). The automated score correlations show a tendency of decreasing as more information from later attempts is considered (from .70 to .66 for biology and from .77 to .71 for psychology).

**Response Time**

Table 6 shows descriptive statistics on the response times of participants in the study. There were no time limits in this study, which presumably resulted in longer response times. The table shows that the second-answer response time is on average 75% to 79% of the first-answer response time. Since about two thirds of the first answers were not awarded full credit, the total average response time for the two attempts was about 50% longer than the average response time for the first answer only (two thirds of 77%). The table also shows that first-answer responses that were not awarded full credit did not take longer to complete than all first-answer responses.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>Md</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First answer</td>
<td>3,589</td>
<td>90</td>
<td>78</td>
<td>66</td>
</tr>
<tr>
<td>First answer—less than full credit</td>
<td>2,341</td>
<td>91</td>
<td>81</td>
<td>64</td>
</tr>
<tr>
<td>Second answer</td>
<td>2,367</td>
<td>71</td>
<td>74</td>
<td>49</td>
</tr>
<tr>
<td>Psychology</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First answer</td>
<td>6,970</td>
<td>73</td>
<td>71</td>
<td>48</td>
</tr>
<tr>
<td>First answer—less than full credit</td>
<td>4,574</td>
<td>75</td>
<td>72</td>
<td>50</td>
</tr>
<tr>
<td>Second answer</td>
<td>4,591</td>
<td>55</td>
<td>65</td>
<td>34</td>
</tr>
</tbody>
</table>

*Note.* Responses longer than 8 minutes (around the 99th percentile) were deleted.
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 fairness of these two formats.

Discussion

The main purpose of this study was to examine the effect of providing immediate feedback and opportunity to revise answers in the context of CR items. In particular, we were interested in whether test-takers are able to improve their scores in response to feedback that indicates their answer is incorrect, and whether scores based on the revised answers prove to be more reliable than scores based on initial answers. Similar to what Attali et al. (2008) found, results of this study showed a significant beneficial effect of providing feedback and allowing subsequent revision of answers. Test-takers were able to increase their scores by about half a standard deviation by using
the second chance of answering the items. Their revised scores were also more reliable than initial scores.

These gains were based on automated scores of subject-area short-answer questions. These scores were less accurate than human scores of the responses. Nevertheless, examinees were able to capitalize on the automated feedback to revise their answers.

The increase in reliabilities can be translated into gains in test length. If we apply the Spearman-Brown formula to the initial score reliabilities (.532 and .590 for biology and psychology, respectively) and predict the amount of lengthening of the test needed to reach the highest score reliabilities (.649 and .650) we get a factor of 1.29 and 1.62. That is, a test based on only one response has to be lengthened by 30% to 60% in order to reach the same reliability of scores based on feedback and one revision. If the increase in response time due to second responses is taken into account (about 50%) this gain in reliability is erased, in contrast to Attali et al. (2008), where even after taking into account the increased response time, the gains in reliability for revised scores were significant. However, it is important to note that simply increasing response times without providing feedback, either by allowing more time or by somehow forcing test-takers to take more time to answer, will likely have a very small effect on reliabilities since this test was not speeded (see Attali, 2005, for an exception with highly speeded number-right MC tests).

This means that the beneficial effect on score reliability cannot be explained by the increased time on the task. A plausible explanation is that the feedback enables a cognitive reinterpretation of the task. With this explanation, the task that the test-taker faces after an initial incorrect response is different from the initial task before the initial answer. Although the item is the same in both cases, the test-taker knowledge is different. In a sense the second attempt forces the test-taker to solve a different problem than she faced before the first attempt. With this interpretation, feedback and the opportunity for revision of answers do really add new items, or problem solving opportunities, to the test.

Furthermore, one can view the provision of feedback and opportunity for answer revision as a minimal “intervention” in the tradition of dynamic assessment and the revised answer as an estimate of the immediate potential for development, the Zone of Proximal Development. With this interpretation the success in revising answers could be measured in addition to the measurement of initial success with the tasks. In traditional assessment, initial success is the only
ability measured. But when examinees are given the chance to revise their answers through cycles of minimal feedback and revision, the revised scores may be seen as a measure of the student sensitivity to feedback and potential for learning. This may be particularly useful in formative assessment or in practice situations.

Another interesting possibility for future research concerns the possible beneficial effect of providing immediate feedback and opportunity to revise answers on test anxiety. In both this study and in Attali et al. (2008), examinees preferred MC tests over open-ended tests, but overwhelmingly preferred receiving feedback and opportunity to revise answers over no feedback, and also indicated that an open-ended test without feedback was the most stressful mode of testing. This pattern of results may reflect the preference of examinees to have some kind of negative feedback on their response, that is, information on the incorrectness of the response. MC tests sometimes provide this kind of information indirectly—when an examinee’s answer is not among the options presented.

In conclusion, advances in automated scoring technology open an intriguing possibility to provide immediate feedback on the correctness of responses and allow examinees to revise their answers. This study found that allowing such interactivity in the testing environment was beneficial to examinees’ scores and contributed to the reliability of scores. These results suggest that such interactivity allows additional cognitive processing of the test tasks that is not usually manifested in a static test and is beneficial to the measurement of test-takers’ ability.
References


for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental

Cognitive Science Society (Ed.), *Program of the twelfth annual conference of the Cognitive
Lawrence Erlbaum.

York: Columbia University, Teachers College.

Vispoel, W. P. (2000). Reviewing and changing answers on computerized fixed-item vocabulary


Wainer, H. (1993). Some practical considerations when converting a linearly administered test to

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 colchicine 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?
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