Research Report

Effect of Immediate Elaborated Feedback on Rater Accuracy

ETS RR–20-09

Yigal Attali

December 2020

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Effect of Immediate Elaborated Feedback on Rater Accuracy

Yigal Attali

Educational Testing Service, Princeton, NJ

Principles of skill acquisition dictate that raters should be provided with frequent feedback about their ratings. However, in current operational practice, raters rarely receive immediate feedback about their scores owing to the prohibitive effort required to generate such feedback. An approach for generating and administering feedback responses to raters is proposed. It consists of automatically designating some responses as feedback responses, sourcing scores, and elaborations for these responses from a group of raters as part of regular scoring and, finally, administering the same responses to all other raters with immediate feedback based on a summary of the available scores and elaborations. This approach allows raters to receive frequent immediate feedback on a regular basis in a sustainable way. In two experimental studies, the effect of frequent immediate feedback (in approximately 25% of responses) on rating accuracy of newly trained raters was investigated. A control condition of no feedback was compared with two types of feedback with elaboration: text explanations of the correct score and a structured form identifying the strengths and weaknesses of the response. Results indicate that feedback had a beneficial effect on rater accuracy and that structured feedback was either equally beneficial to or more beneficial than text explanations.

Keywords Rater training; rater expertise; constructed response assessment

doi:10.1002/ets2.12291

Background

A primary goal of performance assessments is to ensure that raters interpret the scoring guidelines similarly to achieve consistent scores across ratings. Numerous studies of rater behavior have shown that achieving this goal is a continuous challenge (Bachman et al., 1995; Eckes, 2008; Engelhard, 1994; Engelhard & Myford, 2003; Lumley & McNamara, 1995; Weigle, 1998). As a result, the reliability of performance assessments is relatively low when testing time is taken into account (Breland et al., 1987; 1999).

Rater Training and Certification

Because rater variability is such a serious concern for performance assessments, rater training is used to limit such variation (Attali, 2015; Barrett, 2001; Elder et al., 2005; Lumley & McNamara, 1995; Weigle, 1998, 1999). Training of potential raters is increasingly conducted online (Wolfe et al., 2010). It typically includes review and discussion of relevant materials (including the scoring rubric and benchmark responses) and practice with responses that have been assigned consensus ratings and annotations by expert raters. Potential raters are also typically required to pass a certification test at the end of training to qualify for operational scoring (Baldwin et al., 2005; McClellan, 2010).

Operational raters continue their training and development in two ways (McClellan, 2010). First, raters are typically required to pass a *calibration* test (similar to but shorter than the certification test) before every scoring shift. Second, scoring leaders (experienced raters) monitor the quality of raters' ratings in different ways and discuss problems with the raters. Monitoring can be based on backscoring (scoring after the fact) some responses, seeding special *validity* responses (for which an assigned expert rating exists, also called *monitor* responses) among regular operational responses, and examining discrepancies among ratings when operational responses are rated by more than one rater.

However, monitoring of raters rarely results in timely feedback about the score they assigned to a specific student response. This is primarily due to logistical difficulties: In the context of a large-scale assessment with dozens or hundreds of raters, it is very difficult for scoring leaders to backscore a substantial number of ratings for every rater or to develop

Corresponding author: Y. Attali, E-mail: yattali@ets.org

a large enough number of validity responses. For example, in Wolfe et al. (2010), raters scored 400 responses (following training) but received no feedback during scoring. Knoch (2011) provided a summary feedback report to raters 2 weeks after each of several operational administrations, but raters did not receive any feedback during the scoring sessions. As a result, some research has focused on providing summary feedback to raters about their overall severity, consistency, and biases over an extended period of time (Elder et al., 2005; Knoch, 2011; O'Sullivan & Rignall, 2007). Surprisingly, this type of feedback, often based on the output of a many-faceted Rasch analysis, was not generally found to improve rater accuracy (Elder et al., 2005; Knoch, 2011; O'Sullivan & Rignall, 2007).

The purpose of this report is twofold: to propose a feasible and sustainable way (in the context of large-scale assessment) to provide immediate feedback to raters while they are scoring and to report the results of two experimental studies that evaluated the impact of this type of feedback on rater accuracy.

Importance of Feedback

Why is timely feedback to raters important? Newly trained raters cannot be considered experts in applying the scoring rubric for the task they trained on, because in most cases, they have reviewed scores and rationales for just a few dozen responses (Attali, 2019). Therefore the scoring sessions that follow can and should be regarded as a continued opportunity for learning how to apply the scoring rubric to actual responses. In this respect, these scoring sessions can be viewed as formative assessments (Scriven, 1967; Wiliam & Thompson, 2007). For formative assessments, feedback is essential. 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 (Shute, 2008), which can enhance the efficiency of the learning process and result in higher learning outcomes (Mory, 2004). Although feedback in educational contexts may primarily be associated with cognitive processes, it can also operate through affective processes, such as increased effort, motivation, or engagement (Hattie & Timperley, 2007).

Research has shown that the effectiveness of feedback for learning depends on the type of feedback provided (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Shute, 2008; Van der Kleij et al., 2015). Two major distinctions in the literature are the type of feedback and the timing of feedback. Shute (2008) distinguished between various types of feedback based on the specificity and complexity of the feedback: Knowledge of results (KR) only states whether the answer is correct or incorrect; knowledge of correct response (KCR) in addition provides the correct response whenever the answer is incorrect; elaborated feedback (EF) takes different forms, such as an explanation of the correct answer, a worked-out solution, or a reference to study material. Research has suggested that EF (which can take many forms) is the most effective feedback type (Mory, 2004;Shute, 2008; Van der Kleij et al., 2015).

The timing of feedback is another important factor that affects its success (Shute, 2008). Research has found that immediate feedback is generally, but not always, more effective than delayed feedback (Shute, 2008; Van der Kleij et al., 2015), with immediate feedback being more effective for difficult tasks and procedural skills.

A Sustainable Approach for Generating Immediate Feedback

Rating of student responses requires a complex cognitive analysis of responses and their relation to the scoring rubric. As such, the literature on feedback has suggested that immediate feedback will be more effective than delayed feedback for newly trained raters. This literature also recommended that the feedback not be limited to the correctness of the score assigned to a response but that it should elaborate on the reasoning or explanation for this score.

To provide this type of feedback frequently and on a regular basis to a large number of raters would be prohibitive for a small number of scoring leaders. Therefore, in this report, we propose to produce this type of feedback automatically on the basis of scores and elaborations generated by regular raters as part of their regular scoring activities. Furthermore, we propose to base this feedback on a summary of scores and elaborations provided by a group of raters, as an alternative to determining correct scores based on a single scoring leader.

In particular, within a computerized scoring system, the following procedures can be implemented: (a) Regular responses are randomly selected as future feedback responses, (b) a group of raters (e.g., 6 – 10) is assigned to generate feedback for these responses by scoring and elaborating on the reasons for the score, (c) the system automatically generates a consensus (the median) score and elaborations consistent with this score, and (d) the system then assigns the same responses to all other raters to receive immediate feedback on their scores based on the consensus score.

When all raters routinely generate feedback for some responses, these procedures can ensure that all raters receive feedback on a regular basis. For example, assume that the size of the rater pool is 100 and that raters normally assign a single score per response. If no feedback is sourced from regular raters and later presented to other raters, the total number of ratings in the system will be equal to the number of responses (one rating per response). If, on the other hand, feedback responses are used, and for each feedback response, 10 raters generate the feedback for the other 90 raters, then a rate of X% (e.g., 20%) of feedback responses (out of all responses in the system) results in a rate of 90% × X% (18%) of responses for which a rater receives feedback and a rate of 10% × X% (2%) for generating feedback. The cost for the system in terms of excess ratings is X% over a situation where no feedback is provided. Note that feedback responses are efficiently used because all raters either provide or receive feedback for these responses. Note also that to *jump-start* this feedback process for the first few responses, a certain number of feedback responses is needed. These can be obtained by holding out certain responses from the training and certification stage and presenting them to all raters (with immediate feedback) at the initial phases of scoring.

The approach proposed in this report, resulting in raters receiving frequent feedback about their work, can have significant cognitive advantages in developing raters' skills in interpreting the scoring rubric. However, feedback should also be beneficial for increased effort, motivation, and engagement of raters (Hattie & Timperley, 2007) compared to an alternative with no feedback for long stretches of time. Furthermore, generating feedback for other raters as part of this process is likely to have an additional cognitive advantage for raters, one resulting from generating an explanation of their own score. A large research body has demonstrated the effectiveness and generality of self-explanation as an instructional learning strategy (Wylie & Chi, 2014). Self-explanation is a generative learning activity that facilitates deep and robust learning through a process by which students generate inferences and then map these inferences to their existing mental models (Chi, 2000) but also by encouraging students to become cognitively engaged with the learning material (Schworm & Renkl, 2006).

The use of regular raters instead of expert raters for generating feedback could potentially degrade the quality of feedback. On the other hand, the use of a larger number of raters to determine a consensus score for feedback purposes should improve their quality. How this trade-off plays out in practice is an empirical question, but recently, Attali (2019) showed that correct scores used for training and certification tests were of higher quality (as evidenced by the reliability of the certification tests) when generated by a group of regular raters than when generated by a single expert.

Another advantage of consensus scores for feedback is that they can be used to highlight the natural ambiguity about the quality of some responses. Even with the best scoring rubrics, some responses do not fall neatly into one of the levels. In fact, expert raters are known to use this ambiguity to refer to some responses as *low*, *solid*, or *high* within a rating point. Attali et al. (2013) encouraged raters to use this refined scale (effectively tripling the number of rating points), which led to significantly higher score reliability, supporting the validity of these notions. In this context, it is possible that the median score would fall between two score points, and we propose that feedback about this response acknowledge this uncertainty.

Types of Elaboration

Elaboration of scores in the context of rating performance assessments comes in the form of text annotations or explanations of the score in terms of the scoring rubric and the scoring notes. In the empirical studies in this report, several explanations were presented to raters as score elaboration, those that were consistent with the consensus score. We also explored an additional type of elaboration of scores, one that is more structured and therefore would be easier to summarize automatically. To implement the structured feedback approach, a standard form was developed as part of the training materials for each task. The form was based on the scoring rubric and scoring notes for the task and was composed of a list of statements, potential strengths and weaknesses of a response to the task, that could apply to any response (see the appendix for an example). Raters were asked to complete the form (check any applicable points) when they provided (generated) an explanation of their score. Raters were presented with a completed form, based on elaborations of previous raters, when they received feedback on their score. In the studies in this report, a point was checked if at least two-thirds of raters who generated feedback checked that point.

Current Studies

The purpose of the studies in this report was to compare the accuracy of newly trained raters in scoring responses with or without frequent feedback. Three feedback conditions were compared: no feedback, feedback with text explanation elaboration, and feedback with structured form elaboration. Although it is possible, and easier, to provide feedback without

elaboration (KCR), the literature on feedback has clearly suggested that elaboration is important in the context of complex skill acquisition such as rating of performance assessments. Feedback was based on initial expert feedback responses and additional consensus feedback responses. Each study employed a pair of tasks, with more complex tasks in the second study.

Study 1

Method

Materials

Tasks and responses for this study were part of the *CBAL*® learning and assessment tool research initiative to develop assessments that maximize the potential for positive effects on teaching and learning (Bennett et al., 2016). Two CBAL tasks that were developed for middle school students were used in the study. The two tasks were part of two alternate test forms that were constructed to measure the same argumentation skills, but with each form employing a different scenario: whether students should be paid cash for achieving higher grades (cash for grades) and whether advertising to children younger than 12 years of age should be banned (ban ads; Bennett et al., 2016). The particular tasks used in this study asked students to read a "Dear Editor" opinion letter on the topic (cash for grades or ban ads) and then write a note for their classmates in which they identify and explain problems in the letter's reasoning or use of evidence.

Each task had a scoring rubric (in the range of 0 – 4 points), scoring notes, and benchmark responses explaining each rating point. For each task, 525 responses were available for this study. Out of these responses, 50 were selected as training and initial expert feedback responses, and expert-assigned scores, text explanations, and structured feedback were developed for them. The other 475 responses were used in the regular scoring sessions.

Participants

Participants were recruited from Amazon's Mechanical Turk (MTurk) crowdsourcing marketplace. The only qualifications required to participate in the study were U.S. residency, English as a first language, and participation in a previous study of grading student responses conducted by the authors (e.g., Attali, 2015). Apart from this previous study, none of the participants had any substantial experience in grading.

A total of 251 MTurk workers (128 for the ban ads task and 123 for the cash for grades task) completed the training session (in 31 minutes on average) and were paid $6 for their participation in this session. Their ages varied from 19 to 73 years ($M = 34.45$, $Md = 32$, $SD = 10.00$), 49% were women, and most had at least some postsecondary education (12% high school graduates, 22% some college, 14% associate's degree, 44% bachelor's degree, and 8% graduate degree).

Out of the 251 training participants, 122 (49.0%: 47.0% for the ban ads task and 50.0% for the cash for grades task) passed the minimum performance threshold (explained later) and were invited to participate in three additional scoring sessions (with $6 compensation for each of these sessions).

A total of 90 participants completed these additional sessions (45 for the ban ads task and 45 for the cash for grades task) in one of the feedback conditions (four participants started the regular scoring sessions but did not complete them) and are the focus of the analyses.

Training Procedures

Participants in the training session were first asked to read carefully all task materials and then were presented with the 25 training responses in random order. After assigning a score to each of the responses, they received immediate feedback about the correctness of the score they assigned as well as feedback elaboration of either a text explanation of the score (in the control and explanation conditions) or a filled-out structured feedback form (in the structured feedback condition). Every three responses, participants were also asked to elaborate on the score they assigned (before receiving feedback) either with a text explanation of the score (in the control and explanation conditions) or by filling out the structured feedback form (in the structured feedback condition).

In addition, participants received overall feedback about their performance by way of points for each response: 3 points for an exact match between the assigned score and the consensus score and 1 point for a 1-point discrepancy. To pass the

training session successfully, participants had to receive at least 49 points. This threshold could be achieved with 12 exact agreements and 13 adjacent agreements, which roughly corresponds to the minimum level of performance expected in certification tests (Baldwin et al., 2005). Participants were only told they had to accumulate a minimum number of points to proceed to regular scoring but were not told what the threshold was. Following the training session, participants were notified if they had successfully passed training and, if they had passed, were invited for three additional scoring sessions. All scoring was completed within a single day.

Scoring Sessions

The scoring sessions were designed to have approximately the same number of responses and (for the two feedback conditions) around 25% of feedback responses with initial expert feedback responses in the early sessions and consensus feedback responses in the later sessions, with participants generating feedback on other consensus feedback responses in the early stages. Sessions in the control condition were composed of the same responses, except that no feedback was provided for any of the responses. These considerations led to the following design of sessions:

1. Each of the three sessions had approximately 55 responses.
2. There were 30 consensus feedback responses, divided into two sets of 15.
3. There were 25 initial expert feedback responses.
4. There were a total of 445 regular responses, divided into four approximately equal sets.
5. The first session consisted of 28 ± 1 regular responses, 13 initial expert feedback responses, and 15 consensus feedback responses for which the participant generated feedback.
6. The second session consisted of 42 ± 1 regular responses, 6 initial expert feedback responses, and 8 consensus feedback responses on which feedback was given.
7. The third session consisted of 42 ± 1 regular responses, 6 initial expert feedback responses, and 7 consensus feedback responses on which feedback was given.

The composition of sessions is illustrated in Figure 1. In general, in each session, participants could receive feedback (if not in the control condition) on 13–14 responses out of 55 responses (24–25%). The order of responses in each session was randomized. Between 13 and 17 participants completed the study for a given task and feedback condition. Because half of the participants generated feedback for each consensus feedback response (in the first session), the number of ratings the consensus feedback (in the second and third sessions) was based on was between six and nine.

Analyses

Typically, studies that involve performance assessments make an effort to collect two ratings per response to allow the computation of rater agreement statistics to estimate rating quality. In this study, each response was rated by a large number of raters. Therefore analyses of the accuracy of scores were based on absolute discrepancies (distance) between a score and the the average of all ratings provided for a response across the three feedback conditions (termed error). For the 445 regular responses per task, the total number of ratings was 9–13, with a median of 11 ratings (in other words, approximately one-fourth of all 45 participants rated these responses). For the 30 consensus feedback responses per task, the total number of ratings was 45 (in other words, all 45 participants rated these responses). For the 25 initial expert feedback responses per task, the total number of ratings was also 45.

To examine the effect of feedback type on the accuracy of scores, an analysis of covariance was conducted with errors as dependent variable, task and feedback type as independent variables, and errors in training as covariates. For feedback type, two contrasts were tested: feedback (explanation or structured) versus no feedback and explanation feedback versus structured feedback.

Results

Table 1 summarizes the means and standard deviations of errors across tasks and feedback conditions for both the training and scoring phases. No differences in performance during training were found through a 3 (Feedback Type) × 2 (Task) analysis of variance on training phase errors, with all Fs < 1. A moderate correlation was found between training and

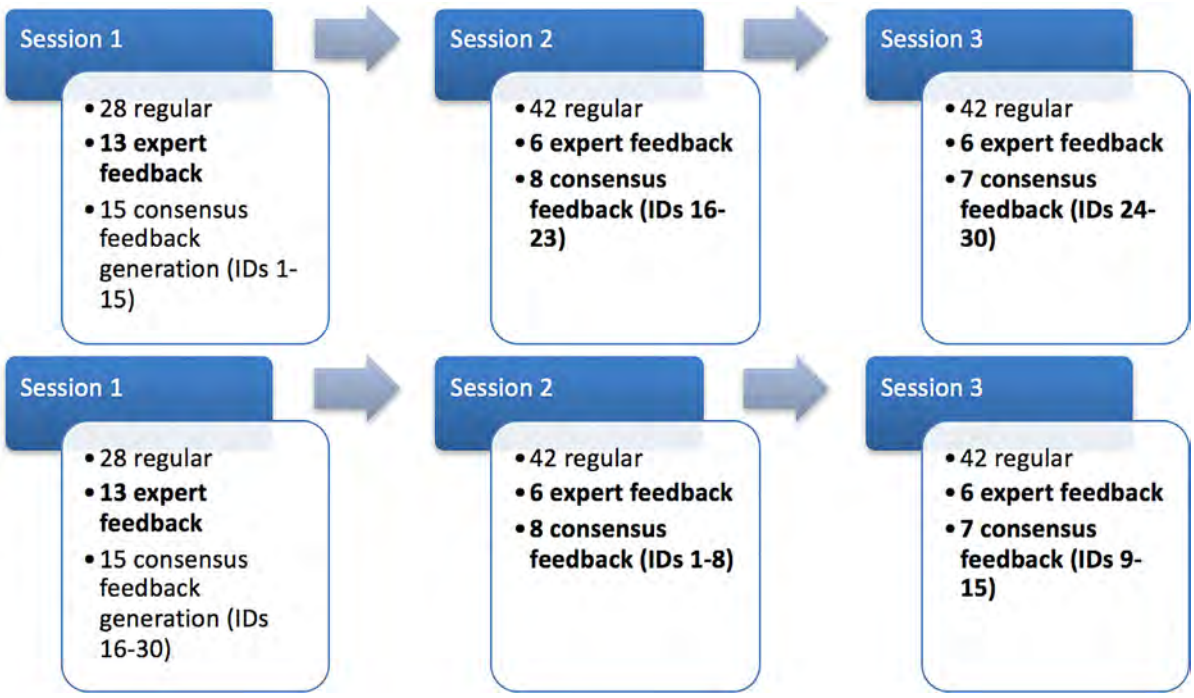

Figure 1 Typical sessions. Responses for which feedback was received are in boldface.

Table 1 Person-Level Summary of Scoring Errors for Study 1

		Training		Scoring	
Feedback	*N*	*M*	*SD*	*M*	*SD*
Ban ads					
None	17	.52	.13	.45	.14
Explanation	14	.51	.17	.42	.07
Structured	14	.52	.14	.43	.07
Cash for grades					
None	17	.55	.14	.58	.17
Explanation	13	.57	.17	.53	.11
Structured	15	.53	.16	.48	.04

Note. Errors defined as absolute distance from consensus scores.

scoring errors ($r = .49$, 95% CI [.31, .63]). Interestingly, the standard deviation of errors in the no-feedback condition was significantly higher than it was in the feedback conditions, as indicated by Brown – Forsythe's test for homogeneity of variance ($F[1, 88] = 12.14$, $p = .001$).

To answer the main research questions, a 3 (Feedback Type) × 2 (Task) analysis of covariance was conducted on scoring phase errors with training phase errors as covariate. The main effect of feedback type did not reach significance ($F[2, 83] = 2.88$, $MSE = 0.01$, $p = .062$, $\hat{\eta}^2_G = .065$). However, planned comparisons for feedback type showed that scoring accuracy was higher (errors were lower) in the feedback conditions (the combined structured feedback and explanation feedback conditions) than in the no-feedback condition ($M = 0.05$, $SE = 0.02$, $p = .022$) but scoring accuracy was not higher with structured feedback than it was with explanation feedback ($M = 0.01$, $SE = 0.03$, $p = .600$). In addition, task ($F[1, 83] = 13.99$, $MSE = 0.01$, $p < .001$, $\hat{\eta}^2_G = .144$) affected scoring errors, but the effect of feedback did not differ by task ($F[2, 83] = 1.24$, $MSE = 0.01$, $p = .294$, $\hat{\eta}^2_G = .029$). Although no interaction between task and feedback type was found, Figure 2 shows that the effects of feedback were stronger for the cash for grades task, which was also the more difficult task.

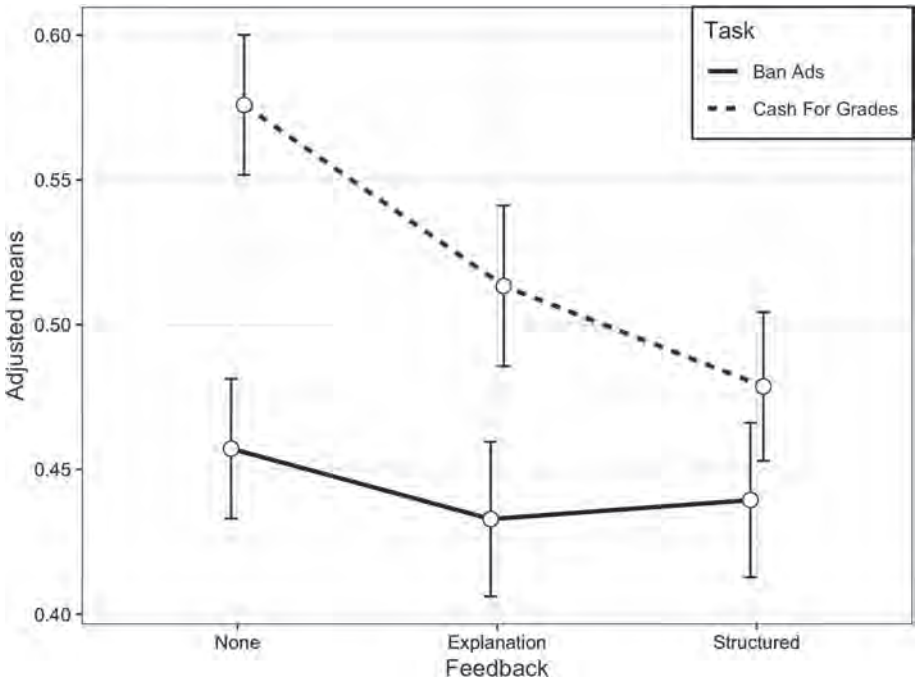

Figure 2 Model predictions for average training errors for Study 1 (with standard error bars).

Study 2

The purpose of Study 2 was to replicate Study 1 with more complex CR tasks and more valid consensus scores. Specifically, more complex essay writing tasks were used in this study, and the consensus scores were based on a large number of ratings produced by an independent group of expert raters. This also allowed us to evaluate the validity of consensus scores based on study participants and to compare the results of using different definitions of consensus scores.

Method

Materials

Materials for this study were based on the analytical writing measure of a large-scale, college-level standardized assessment (which also includes a measure of reading comprehension and verbal reasoning ability). This assessment comprises two essay writing tasks. In the issue task, the test taker is asked to discuss and express his or her perspective on a topic of general interest. In the argument task, a brief passage is presented, from which the author makes a case for some course of action or interpretation of events by presenting claims backed by reason and evidence. The test taker's task is to discuss the logical soundness of the author's case by critically examining the line of reasoning and the use of evidence.

For this study, two argument tasks (labeled "candidates" and "waste") and operational essays written in response to tasks were used. In addition, 16 ratings from experienced raters were collected in a previous study (Attali et al., 2013) for each of the essays.

Each task had a scoring rubric (in the range of 0–6 points), scoring notes, and benchmark responses explaining each rating point. For each task, 123 responses were available for this study. Out of these responses, 23 were selected as training and initial expert feedback responses, and expert assigned scores, text explanations, and structured feedback were developed for them. The other 100 responses were used in the regular scoring sessions.

Participants

Participants for this study were recruited from MTurk with the same qualifications as in Study 1. A total of 258 MTurk workers (142 for the candidates task and 116 for the waste task) completed the training session (in 37 minutes on average) and were paid $7 for their participation in this session. Their ages varied from 19 to 69 years ($M = 35.69$, $SD = 10.25$),

45% were women, and most had at least some postsecondary education (11% high school graduates, 19% some college, 16% associate's degree, 44% bachelor's degree, and 10% graduate degree).

Out of the 258 training participants, 133 (52.0%: 44.0% for the candidates task and 60.0% for the waste task) passed the minimum performance threshold (explained later) and were invited to participate in three additional scoring sessions (with $7 compensation for each of these sessions).

A total of 80 participants completed these additional sessions (39 for the candidates task and 41 for the waste task) in one of the feedback conditions and are the focus of the analyses.

Training Procedures

This study employed the same training procedures as Study 1, with the exception that participants had 15 training responses and were required to accumulate at least 29 (out of 45) points to pass training.

Scoring Sessions

This study employed the same general scoring procedures as Study 1, with the following design of sessions:

1. Each of the three sessions had approximately 55 responses.
2. There were 20 consensus feedback responses, divided into two sets of 10.
3. There were eight initial expert feedback responses.
4. There were a total of 80 regular responses, divided into two equal sets.
5. The first session consisted of 6 – 7 regular responses, 6 initial expert feedback responses, and 10 consensus feedback responses on which the participant provided feedback.
6. The second session consisted of 16 – 17 regular responses, 1 initial expert feedback response, and 5 consensus feedback responses on which feedback was given.
7. The third session consisted of 16 – 17 regular responses, 1 initial expert feedback response, and 5 consensus feedback responses on which feedback was given.

In general, in each session, participants could receive feedback (if not in the control condition) on 6 responses out of 22 – 23 responses (26% – 27%). The order of responses in each session was randomized. Between 13 and 17 participants completed the study for a given task and feedback condition. Because half of the participants provided feedback on each consensus feedback response (in the first session), the number of ratings on which the consensus feedback (in the second and third sessions) was based was between six and nine.

Analyses

Similar analyses were conducted for this study as for Study 1, except that consensus scores were calculated on the basis of 16 expert judgments available for each response.

Results

Table 2 summarizes the means and standard deviations of errors across tasks and feedback conditions for both the training and scoring phases. No differences in performance during training were found through a 3 (Feedback Type) × 2 (Task) ANOVA on training phase errors, with all Fs < 1. A weaker correlation was found between training and scoring errors ($r = .20$, 95% CI [−.02, .41]) than in Study 1, possibly as a result of the lower number of training responses. As in Study 1, the standard deviation of errors in the no-feedback condition was higher than it was in the feedback conditions, but the difference was not significant ($F[1, 78] = 2.47$, $p = .120$).

To answer the main research questions, a 3 (Feedback Type) × 2 (Task) analysis of covariance was conducted on scoring phase errors with training phase errors as covariates. As Figure 3 shows, the main effect of feedback type was on scoring errors ($F[2, 73] = 7.52$, $MSE = 0.02$, $p = .001$, $\hat{\eta}^2_G = .171$). In this study, too, planned comparisons showed that scoring accuracy was higher (errors were lower) in the feedback conditions (the combined structured feedback and explanation feedback conditions) than in the no-feedback condition ($M = 0.11$, $SE = 0.03$, $p = .001$), but in this study, scoring accuracy

Table 2 Person-Level Summary of Scoring Errors for Study 2

		Training		Scoring	
Feedback	*N*	*M*	*SD*	*M*	*SD*
Candidates					
None	12	.48	.14	.75	.18
Explanation	14	.47	.15	.62	.17
Structured	13	.57	.08	.56	.12
Waste					
None	16	.59	.16	.65	.17
Explanation	12	.54	.21	.63	.09
Structured	13	.59	.15	.57	.14

Note. Errors defined as absolute distance from consensus scores.

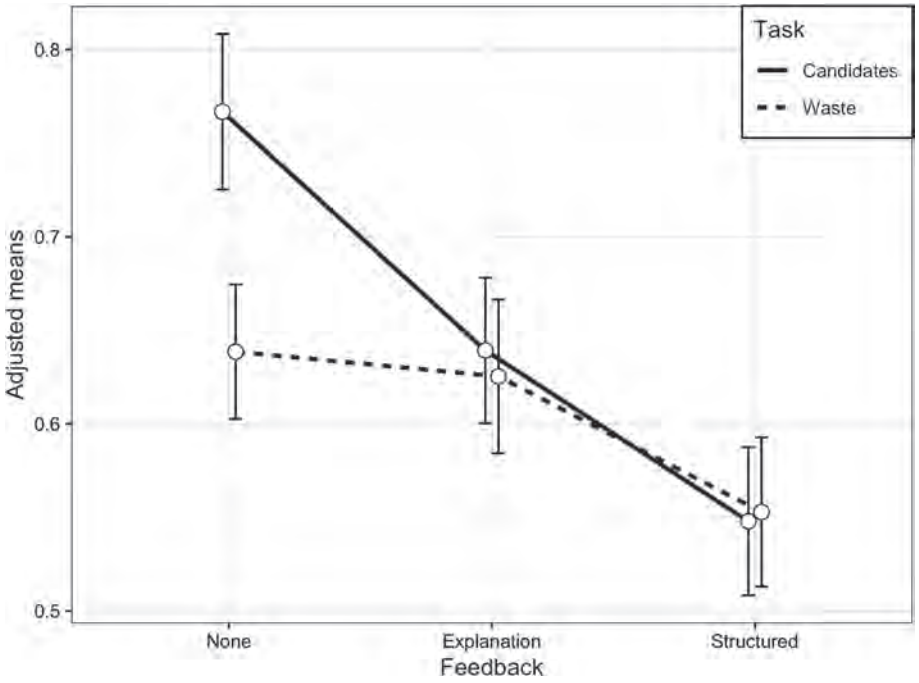

Figure 3 Model predictions for average training errors for Study 2 (with standard error bars).

was also higher with structured feedback than it was with explanation feedback ($M = 0.08$, $SE = 0.04$, $p = .046$). In this study, task did not affect errors ($F[1, 73] = 1.93$, $MSE = 0.02$, $p = .169$) and the effect of feedback did not differ by task ($F[2, 73] = 1.71$, $MSE = 0.02$, $p = .188$).

In addition, a parallel ANCOVA was conducted on the basis of consensus scores calculated from study participants (similarly to Study 1) instead of independent expert ratings with identical results (the consensus scores were highly correlated [$r = .87$]).

Discussion

Feedback Findings

This report investigated the effect of frequent immediate feedback (in approximately 25% of responses) on the accuracy of newly trained raters of student responses to performance assessments. Across two studies and four tasks, a control condition of no feedback was compared with two types of KCR feedback with elaboration: text explanations of the correct score and a structured form identifying the strengths and weaknesses of the response.

A large body of literature on the effects of feedback in testlike situations has suggested that feedback helps learners determine performance expectations, judge their level of understanding, become aware of misconceptions, and provide clues about how to correct mistakes and improve performance (Hattie & Timperley, 2007; Shute, 2008; Van der Kleij et al., 2015), which can enhance the efficiency of the learning process and result in higher learning outcomes (Mory, 2004). Our findings are consistent with this literature. In both studies, the feedback conditions led to higher levels of rater accuracy compared to the no-feedback condition. The *d* effect sizes in Study 1 were 0.23 and 0.66 for the ban ads and cash for grades tasks, respectively, and in Study 2, they were 1.03 and 0.41 for the candidates and waste tasks, respectively.

A related interesting finding is the reduced variability in accuracy across raters who received feedback compared with variability across raters who did not receive feedback. The standard deviations of error across raters in the feedback conditions were 52% and 76% in the no-feedback condition in Studies 1 and 2, respectively (although this reduction was not statistically significant in Study 2). Therefore it seems that part of the effect of feedback is to help level the performance of raters within a task.

Finally, task difficulty may moderate the beneficial effect of feedback. The more difficult the task is for raters (without feedback), the greater feedback seemed to benefit raters across the studies. For example, the ban ads task was the easiest (lowest levels of errors) for raters with no feedback and showed the smallest effect for feedback, whereas the candidates task was hardest for raters with no feedback and showed the largest effect for feedback. This moderating effect of task difficulty should be explored more thoroughly in future research, but it makes intuitive sense — when the rubric is easy to apply, less rater training is required. To summarize, future research should explore the possibility that feedback levels the performance of raters both *within* and *between* tasks.

Type of Elaboration

Two types of feedback elaboration were explored in this report. In addition to traditional text explanations of the score response, an alternative elaboration of scores was implemented through a task-specific form. The form was based on the scoring rubric and scoring notes for the task and was composed of a list of statements, potential strengths and weaknesses of a response to the task, that could apply to any response. This structured elaboration may provide more pertinent information to raters about a particular response and is easier to summarize automatically when multiple raters generate elaborations (by counting the number of elaborations that checked each item in the form). Note that KCR feedback was provided in both types of elaboration. Therefore these two feedback conditions shared an important component of feedback. A comparison of the two elaboration types found no statistical difference in Study 1, but structured feedback led to more accurate scoring in Study 2. Although more research is required to understand how elaborations help raters, one possibility is that task difficulty or rubric complexity affects the relative benefit of different types of elaborations, with more difficult or complex tasks benefiting from a structured elaboration. From a practical point of view, it is possible that a combination of both structured and free text elaboration could be most helpful to raters.

Limitations

One limitation of the studies in this report is the focus on inexperienced raters. It is possible that the effects of feedback would dissipate with more experienced raters. Presumably, experience would make the tasks easier for raters, which could result in a reduced need for feedback. Experience, however, does not automatically translate to expertise, and feedback is an important mechanism for helping learners capitalize on their experience for skill acquisition. The literature on rater training and experience effects has supported the idea that experience in itself is not a determinant of rating quality, which is driven more by training and feedback (Attali, 2015; Lim, 2011; Weigle, 1998).

Another related limitation of the studies in this report is the focus on short-term effects with just a few hours of rating. It is possible that over extended periods of time, feedback would have different effects. One possibility, consistent with the idea that experience renders feedback unnecessary, is that over time, the effects of feedback would dissipate. It is, however, entirely possible that feedback effects are amplified over time, as raters with no feedback gradually drift into their own idiosyncratic interpretation of the rubric.

Finally, although the studies in this report were based on two types of tasks and two tasks within each type, more research is needed to generalize the results across a wide range of tasks.

Implications for Practice

In current practice, monitoring raters is performed by scoring leaders. This leads to feedback being provided infrequently and with an emphasis on summative feedback or, if feedback relates to specific responses, with a significant delay. In this report, we proposed to generate immediate feedback to raters by automatically designating some responses as feedback responses; sourcing scores and elaborations for these responses from a larger group of raters; and finally, administering the same responses to all other raters with immediate feedback based on a summary of the available scores and elaborations. This approach allows raters to receive frequent immediate feedback on a regular basis in a sustainable way. It also uses feedback responses in the most efficient way, because all raters either produce or receive feedback on these responses. Moreover, the same responses that are used to provide feedback to raters can readily be used to monitor the overall quality of raters. It is also important to note that generating frequent immediate feedback to raters does not preclude the use of longer term summative feedback to raters, including discussions with scoring leaders about overall performance.

Although in the studies reported here, feedback was generated by regular raters, the approach can be extended to include scoring leaders as feedback generators. In other words, scoring leaders together with regular raters can generate feedback for future use. Furthermore, although consensus scores are generated automatically, the system could differentially weight the scores and elaborations of different contributors, in particular, weighting more heavily contributions from scoring leaders and other expert raters.

An important practical aspect of providing frequent feedback to raters is cost. As was shown, providing immediate feedback at a particular rate (out of all responses) would require approximately the same rate in excess ratings over a situation where no feedback is generated and provided. In the studies reported here, feedback was provided approximately 25% of the time. In practice, as raters gain experience, it may be reasonable to reduce the rate of feedback responses, although from both an affective and cognitive perspective, some feedback may be beneficial even for experienced raters.

Conclusions

A large body of research has suggested that frequent immediate elaborated feedback to raters would support expertise development and increase engagement. In line with this research, we found that raters receiving frequent immediate feedback (approximately 25% of the time) scored more accurately than raters who did not receive feedback on their scores. To allow raters in operational large-scale settings to reap the benefits of frequent immediate feedback, we propose to produce this type of feedback automatically on the basis of scores and elaborations generated by a group of regular raters as part of their regular scoring activities.

References

Attali, Y. (2015). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99 – 115. https://doi.org/10.1177/0265532215582283

Attali, Y. (2019). Rater certification tests: A psychometric approach. *Educational Measurement: Issues and Practice*, *38*(2), 6 – 13. https://doi.org/10.1111/emip.12248

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*(1), 125 – 141. https://doi.org/10.1177/0265532212452396

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238 – 257. https://doi.org/10.1177/026553229501200206

Baldwin, D., Fowles, M., & Livingston, S. A. (2005). *Guidelines for constructed-response and other performance assessments 2005*. Educational Testing Service.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, *2*(1), 49 – 58.

Bennett, R. E., Deane, P., & van Rijn, P. (2016). From cognitive-domain theory to assessment practice. *Educational Psychologist*, *51*(1), 82 – 107. https://doi.org/10.1080/00461520.2016.1141683

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework* (College Board Report Series No. 99-3). College Entrance Examination Board. https://doi.org/10.1002/j.2333-8504.1999.tb01801.x

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill* (Research Monograph No. 11). College Entrance Examination Board.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 5, pp. 161 – 238). Lawrence Erlbaum.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Elder, C., Knoch, U., Barkhuizen, G., & van Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly: An International Journal*, *2*(3), 175–196. https://doi.org/10.1207/s15434311laq0203_1

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Engelhard, G., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English literature and composition program with a many-faceted Rasch model* (Research Report No. RR-03-01). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01893.x

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior—A longitudinal study. *Language Testing*, *28*(2), 179–200. https://doi.org/10.1177/0265532210384252

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543–560. https://doi.org/10.1177/0265532211406422

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54–71. https://doi.org/10.1177/026553229501200104

McClellan, C. A. (2010). *Constructed-response scoring—doing it right* (R&D Connections No. 13). Educational Testing Service.

Mory, E. H. (2004). Feedback research revisited. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745–783). Lawrence Erlbaum.

O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 446–478). Cambridge University Press.

Schworm, S., & Renkl, A. (2006). Computer-supported example-based learning: When instructional explanations reduce self-explanations. *Computers and Education*, *46*(4), 426–445. https://doi.org/10.1016/j.compedu.2004.08.011

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Rand McNally.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, *85*(4), 475–511. https://doi.org/10.3102/0034654314564881

Weigle, S. C. (1998). Using facets to model rater training effects. *Language Testing*, *15*(2), 263–287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), 145–178. https://doi.org/10.1016/S1075-2935(00)00010-6

Wiliam, D., & Thompson, M. (2007). Integrating assessment with learning: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Lawrence Erlbaum. https://doi.org/10.4324/9781315086545-3

Wolfe, E. W., Matthews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *Journal of Technology, Learning, and Assessment*, *10*(1), 4–21.

Wylie, R., & Chi, M. T. H. (2014). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 413–432). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.021

Appendix

The structured feedback form for the cash for grades task is presented in the following nested list. Each of the "leaf" (lowest level) items in the list (14 in this case) could be checked to describe a response.

- Response correctly identified problems related to
  - Daughter's past grades
    - The writer's daughter used to do well in school without a cash incentive, so cash may not be the reason she gets good grades now.

        - Perhaps the writer's daughter did not do well last year for social or personal reasons, not because she wasn't receiving cash rewards.
    - New versus old school
        - Perhaps the daughter's new school has a more relaxed grading policy than her old school, which would make it easier for her to do well.
        - Perhaps the small class sizes and good facilities at the new school, rather than the cash rewards, are enabling her to earn good grades.
        - It's also possible that simply moving to a new place has changed the daughter's attitude and motivated her to work harder.
    - Other reasoning issues
        - In the first paragraph, the writer objects to giving cash rewards because kids might waste the money, but in the third paragraph, he says he lets his daughter spend her money however she wants.
        - The writer mentions his son in the first paragraph but does not refer to him again. Did his grades go up at the new school, too?
        - The writer generalizes about cash rewards based on only one example: his daughter.
- Problems with response
    - Misinterprets or distorts the letter
    - Includes irrelevant information
    - Misinterprets or distorts the writing task (e.g., refers to problems with writing features, not reasoning)
    - Identifies only generic reasoning problems that could apply to almost any argument, not this specific argument (e.g., refers to the need for "more evidence" or "better examples")
    - Off topic
    - Only random key strokes

Suggested citation:

Attali, Y. (2020). *Effect of immediate elaborated feedback on rater accuracy* (Research Report No. RR-20-09). Educational Testing Service. https://doi.org/10.1002/ets2.12291

Action Editor: Marna Golub-Smith

Reviewers: Bridgid Finn and Peter van Rijn