Measuring the Power of Learning.®

Research Memorandum
ETS RM–19-13

Examining the Perceived Effectiveness of Rater Feedback

Cathy Wendler

Dawn Leusner

Valerie Thompson

Florencia Tolentino

October 2019

ETS Research Memorandum 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 Memorandum 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 Memorandum 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.

Examining the Perceived Effectiveness of Rater Feedback

Cathy Wendler, Dawn Leusner, Valerie Thompson, and Florencia Tolentino
Educational Testing Service, Princeton, New Jersey

October 2019

Corresponding author: C. Wendler, E-mail: wendlercathy@gmail.com

Suggested citation: Wendler, C., Leusner, D., Thompson, V., & Tolentino, F. (2019). *Examining the perceived effectiveness of rater feedback* (Research Memorandum No. RM-19-13). Princeton, NJ: Educational Testing Service.

Action Editor: John Mazzeo

Reviewers: Bridgid Finn and Edward Wolfe

Abstract

This study examined rater perceptions of the effectiveness of feedback practices used by testing programs at Educational Testing Service. Practices used in rater training and during scoring were examined. The study involved conducting one-on-one telephone surveys with trained and experienced raters. A total of 36 raters were surveyed, with 17 being raters for the English Language Proficiency Assessments for California, 10 for the *GRE*® General test, and 9 for the *TOEFL iBT*® test. Survey questions covered 4 categories: (a) feedback practices used during training and calibration, (b) feedback practices used during operational scoring, (c) information received from a scoring leader, and (d) information specific to the performance of the individual rater. Results indicate that the level, type, and frequency of feedback appear to define its usefulness to raters. To be useful, feedback on scoring accuracy needs to be immediate and concise and to provide specific information that indicates why a rater's assigned score was incorrect. In addition, feedback on scoring rate needs to be provided in a context that it is easily interpretable and understandable by raters. Feedback from scoring leaders is perceived as valuable, regardless of how it is provided to raters. Raters with less experience desire feedback more frequently than those with more experience. Finally, the method of providing the feedback must be easily accessible while in the scoring system—either displayed on screen or easily obtained through a link.

Key words*:* CR scoring, human raters, feedback

Acknowledgments

The authors acknowledge and thank the following individuals for their work and input on this project: Heather Fell, Andrea Napoli, Gail Strelko, Jane Vaughn, and Edward Wolfe.

Table of Contents

List of Tables

Page

Many constructed-response (CR) tasks and item types, such as essays, speech samples, and short answers, lend themselves to automated scoring. However, not all CR tasks or items can be scored using automated scoring, and thus human raters will likely remain a critical part of CR scoring in the near future. In addition, evaluations of automated scoring models are frequently based on comparisons with human scoring results, which assume that human raters produce accurate and reliable ratings. Challenges associated with scoring CRs are not new (see Bejar, 2017), and one challenge in using human raters is ensuring that the scores produced by the raters remain consistent, appropriately reflect the scoring rubric, and do not become less accurate within or across scoring sessions.

Three high-level phases must be considered when using human raters: (a) Raters must be trained prior to engaging in operational scoring work; (b) they must be qualified prior to operational scoring by demonstrating their ability to score reliability at an acceptable level of accuracy; and (c) rater performance must be monitored during operational scoring to ensure continued scoring accuracy. The variability associated with human raters refers to the measurement error related to raters (Engelhard, 2002; Wolfe, 2014), and rater variability may impact the reliability of scores (Braun, 1988). For example, Cason and Cason (1984) showed that, if not properly managed, rater errors can explain as much as or more score variability as test taker ability. Therefore adequate training, qualification, and monitoring of raters are crucial elements in ensuring valid and accurate scores. An inherent part of these three phases is providing raters with information that allows them to modify their behavior as needed.

As part of operational scoring, raters must use information and knowledge that is acquired during the training and qualification phases. While few empirical studies have been done on how raters "learn" and apply what they have learned, theories from cognitive and learning sciences are useful in providing guidance on methods to help promote and sustain learning in the rating context. First, the concept of long-term retention of knowledge is an important notion in learning. Studies have shown that with longer intervals over which knowledge needs to be retained, the loss of knowledge is greater, and that gaps in using that knowledge increase the loss, especially for tasks that require decision-making or problem solving (see Arthur, Bennett, Stanush, & McNelly, 1998).

Retention and appropriate application of information by raters are critical in the context of operational scoring, especially because many raters score during long scoring periods

followed by no scoring for days, weeks, or even months. Finn, Wendler, and Arslan (2018) and Finn, Wendler, Ricker-Pedley, and Arslan (2018) examined the impact of the number of days in a scoring gap (i.e., nonconsecutive scoring days) and found that gaps in the number of scoring days were associated with a decrease in scoring accuracy, thus providing support for the decline in skills discussed in Arthur et al. (1998).

Another concept from cognitive and learning sciences is goal-setting theory. This theory posits that setting expectations regarding desired performance levels increases motivation levels and performance for students over just telling students to "do their best" (Locke & Latham, 1990). Goal-setting theory has been shown to increase achievement and performance in a number of settings, including educational and employment environments (Latham & Locke, 2007; Locke & Latham, 1990; Morisano, Hirsh, Peterson, Pihl, & Shore, 2010; Wiese & Freund, 2005). Applied to rater behavior, this theory suggests that raters who are given a particular scoring goal (both scoring rate and accuracy) would be more likely to score at that level.

Wendler, Glazer, and Bridgeman (2018) examined whether setting scoring rate expectations would influence the ability of a rater to score at a particular pace and if this expectation would differentially impact accuracy levels for raters who typically read at a slow, medium, or fast read rate. Results indicate no significant differences in scoring rates for raters who were given expected scoring rates compared to those who set their own scoring pace. However, slow and medium raters in both conditions were able to increase their scoring rate (i.e., score more essays in an hour) with no significant effects on rater accuracy.

A third important concept from cognitive and learning sciences is feedback. Feedback can be both positive and negative; feedback can provide information on what individuals are doing correctly and what they are doing incorrectly. In many ways, feedback is a form of behavioral reinforcement. Classical studies on the impact of providing reinforcement (e.g., Skinner, 1953) have indicated that providing both positive and negative information is useful in modifying behavior. Bandura (1977) agreed with the role of reinforcement in learning and further postulated that human behavior can be controlled through continuous interaction between cognitive, behavioral, and environmental reinforcement.

The concept of positive and negative reinforcement—or feedback—is viewed as a mechanism for both motivating learning and improving the acquisition of skills and knowledge (see Shute, 2008), and the role of feedback in learning has been examined for at least 5 decades.

Providing information about the correctness of an answer appears to impact learning and retention to some degree. For example, Pashler, Cepeda, Wixted, and Rohrer (2005) found that providing the correct answer to subjects after an incorrect response was given during an associative learning task improved performance during the initial learning session and also increased final retention of the material.

However, while providing appropriate information at an appropriate level does seem to have an impact on skill acquisition and retention, providing learners with specific, appropriate feedback about the correct answer appears to be more effective. Finn, Thomas, and Rawson (2018) examined whether providing elaborated feedback to participants would increase conceptual understanding and boost knowledge. During the course of two experiments, they provided some participants with feedback that included an example and some with feedback that only included the correct answer. Results indicate that providing examples as part of feedback when learning about educational concepts increased performance on previously tested items and on new, related items.

Wolfe, Winchester, and Rupp (2018) discussed research on feedback as it relates to using human raters in essay scoring. They summarized previous research along four lines: (a) the type and purpose of the feedback, (b) the direction (positive vs. negative) of the feedback, (c) the level of complexity and content of the feedback, and (d) the timing and delivery of the feedback. They concluded that raters tend to overcompensate when they are provided feedback on their performance (e.g., raters who score too leniently began to score too severely following feedback, and vice versa); however, this overcompensation can be mitigated depending on the type of feedback presented. They suggested that, at a minimum, raters should be provided the correct response with the least complex explanation of why the response is correct.

The concepts of knowledge decline, setting expectations, and providing feedback are all relevant to understanding, monitoring, and modifying rater performance. In the current study, we examine how individuals who currently act as raters perceive the effectiveness and limitations of some of the feedback practices currently used or proposed by Educational Testing Service (ETS) as part of the scoring process and how these practices may impact both scoring accuracy and scoring rate.

Study Design and Method

The study involved conducting telephone surveys with trained and experienced raters. A total of 36 raters were surveyed, with 17 being raters for the English Language Proficiency Assessments for California (ELPAC), 10 for the *GRE*® General test, and 9 for the *TOEFL iBT*® test (called here the *TOEFL*® test). All raters only acted as raters for ETS, but about 52% of the raters had experience with scoring multiple testing programs. Two ETS staff conducted the surveys. Both interviewers were first trained using a formal script that was followed during the one-on-one telephone sessions.

The surveys took place over a 7-week period. Raters were told that ETS wanted their input on the usefulness of various features that were hoped to help raters as part of the scoring process, and in particular, ETS wanted input as to information that might be helpful as they operationally scored. They were assured that anything they said would be kept confidential. Raters were compensated at their usual rate of pay.

Survey questions were categorized into four sections that reflected the type of feedback: (a) given during training and calibration, (b) given during operational scoring, (c) received from a scoring leader, and (d) specific to the performance of the individual rater. The appendix displays the survey questions and options.

The number and percentage of raters responding to each survey question option were computed for the overall group and by program. In addition, open-ended responses were analyzed using NVivo12, which automatically groups information from the responses based on thematic, high-level categories. The high-level categories were defined by the first author and confirmed by the second author. The frequencies that come from such an analysis help in understanding the raters' experiences or views but do not indicate the relative importance of each view.

Results

Training and Calibration

Raters were asked about their experience with and perception of two types of feedback features they might have encountered as part of training. The first, the chat feature, allows raters to receive feedback by communicating, interacting, and exchanging messages with a scoring leader over the Internet. Scoring leaders are experts who oversee, monitor, and mentor raters; they may be external consultants who are experienced raters themselves or ETS content

specialists. Raters use the chat function while in the ETS proprietary online scoring system (Online Network for Evaluation, or ONE), which enables raters to score responses to many types of CR tasks, written or spoken, via secure Internet access. While in ONE, raters and scoring leaders can use the chat feature to interactively discuss scoring questions or other concerns.

The second feature, annotated feedback, is an automatic function in ONE where raters are given an explanation about why a particular response depicts a particular score. Annotated feedback is generic in nature and is created by ETS content specialists. Annotated feedback only focuses on the correct score for a response, and thus, it is only provided to raters if they assign an incorrect score to a response during training. Table 1 presents the number and percentage of raters responding to each survey question option for both the chat and annotated feedback features.

Table 1. Number and Percentage of Raters Responding to Each Survey Question Option on Training and Calibration Feedback Features

Survey question	Response option	All raters		ELPAC raters		GRE raters		TOEFL raters	
		No.	%	No.	%	No.	%	No.	%
Chat feature									
Frequency during training	Frequently	3	8	2	12	0	0	1	11
	Occasionally	12	33	7	41	2	20	3	33
	Never	21	58	8	47	8	80	5	56
If never, would it be helpful?	Yes	19	90	6	75	8	100	5	100
	No	1	5	1	13	0	0	0	0
Would chat be helpful during calibration?	Yes	24	67	15	88	5	50	4	44
	No	11	31	1	6	5	50	5	56
Annotated feedback									
Frequency during training	Frequently	12	33	5	29	4	40	3	33
	Occasionally	17	47	9	53	3	30	5	56
	Never	6	17	3	18	2	20	1	11
If never, would it be helpful?	Yes	5	83	2	67	2	100	1	100
	No	1	17	1	33	0	0	0	0
Was the frequency appropriate?	Too little	4	11	0	0	3	30	1	11
	Too much	0	0	0	0	0	0	0	0
	Just right	25	69	14	82	4	40	7	78
Would annotated feedback be helpful during calibration?	Yes	28	78	14	82	8	80	6	67
	No	8	22	3	18	2	20	3	33

Note. Totals may not add to 100% due to nonresponses or rounding. ELPAC = English Language Proficiency Assessments for California.

Overall, the majority of raters (58%) indicated they never used the chat feature during training; this was most pronounced for GRE raters, where 80% indicated they never used chat. Roughly one half of the ELPAC and TOEFL raters indicated they had used the chat feature occasionally, with the other half indicating they never used chat. However, most raters (90%) who did not use chat during training indicated that it would have been helpful to have done so. They indicated that using chat would have provided immediate (faster and quicker) feedback from the scoring leader and helped him or her become more accurate (e.g., "Having more accessibility to someone instantaneously when you're lost or confused would be helpful. It would be a more pleasant experience to be able to communicate with someone"; "You can get fast responses to questions you may have"). However, one rater felt chat should not be available during training, indicating that "Training is self-explanatory."

When asked if they felt that having chat available during calibration would be helpful, about two thirds (67%) felt it would. A higher percentage of ELPAC raters (88%) felt that it would be helpful compared to GRE (50%) and TOEFL (44%) raters. This difference may be the result of the level of rater experience. GRE and TOEFL raters are, in general, more experienced than ELPAC raters and therefore are very familiar with the role of calibration. Several raters felt the availability of chat during calibration would help them understand where they were scoring inaccurately (e.g., "Sometimes some of the responses can be a little tricky"; "Getting feedback about where you made mistakes and potential reasons"; "There are sometimes nuances that you might need some clarification on"). But a number of raters also felt that having chat available during calibration would degrade the reason for calibration (e.g., "Probably not because there is nothing a [scoring leader] can do to help us get through calibration so that would be superfluous"; "[It] defeats the purpose of calibration, which is for the rater to complete on their own"; "Not really, calibration means concentration and the chat feature could be distracting"). A few raters indicated chat could serve a purpose following calibration (e.g., "It would be helpful if you fail calibration and could discuss the results in a little more detail").

A slightly different picture emerged when asked about annotated feedback during training. In this case, just under one half (47%) of the raters indicated that they had occasionally received annotated feedback during training, and one third (33%) received it frequently. Results across ELPAC and TOEFL raters are similar; however, more (40%) GRE raters reported receiving annotated feedback frequently compared to those receiving it occasionally (30%) or

never (20%). The majority of raters (69% overall, 82% ELPAC, 40% GRE, and 78% TOEFL) felt the frequency of receiving annotated feedback was just right. For those few raters who indicated they never received annotated feedback during training, all except one (an ELPAC rater) felt it would have been helpful.

When asked if they felt that receiving annotated feedback during calibration would be helpful, the majority (78%) of raters felt it would be. They believed that knowing which samples they missed and getting feedback that provided a rationale for the correct answer would help them score more accurately (e.g., "It kind of lets you know the reasoning behind why the essay was given the score it was given so you know how to score the essays moving forward"; "The calibration process is pretty opaque. All you do is get a score at the end and you have no idea which responses were incorrect and why or how you can do better on the next calibration if you failed it"; "Yes, if it better prepares you to score and be accurate"). However, many raters felt that having detailed feedback at the end of calibration, not during calibration, would be the most beneficial (e.g., "At the end of calibration it would be helpful to see the true score and why"; "If [you do] not pass calibration you don't know what you did wrong"; "It would nice to have it after I submit my calibration to gain from it. It's frustrating when I fail and I don't know which response I scored off and why"). Other raters felt it would not be helpful and indicated that receiving annotated feedback during calibration would result in inefficiencies (e.g., "If I have to stop and read something to get an explanation, that's great, but that stops my momentum and my thought process"; "It will make the process longer; it would take more time") or should not be necessary (e.g., "By the time you're calibrating you really need to know what you're doing. That's what calibration is all about"; "Enough information is provided. If that kind of feedback is allowed, the scorer does not become comfortable scoring independently. Working it out is a better advantage for having confident and capable scorers, as well as for efficiency. You're given everything you need, one more thing is not necessary").

Operational Scoring

Raters were then asked about their experience with and perception of the chat feature and annotated feedback as part of operational scoring. While the chat feature worked the same way during scoring as it did during training, annotated feedback is accessed through a link called "My Feedback" that appears in the ONE system. Table 2 presents the number and percentage of raters responding to each survey question option.

Table 2. Number and Percentage of Raters Responding to Each Survey Question Option on Scoring Feedback Features

Survey question	Response option	All raters		ELPAC raters		GRE raters		TOEFL raters	
		No.	%	No.	%	No.	%	No.	%
Chat feature									
Frequency during scoring	Frequently	11	31	3	18	4	40	4	44
	Occasionally	25	69	14	82	6	60	5	56
	Never	0	0	0	0	0	0	0	0
Who initiated chat?	Rater	8	22	4	24	2	20	2	22
	Scoring leader	7	19	4	24	2	20	1	11
	Both	21	53	9	53	6	60	6	67
Did using chat affect scoring accuracy?	Yes	25	69	13	76	8	80	4	44
	No	10	28	4	24	2	20	4	44
Did using chat affect scoring rate?	Yes	21	58	11	65	4	40	6	67
	No	15	42	6	35	6	60	3	33
Communicate with scoring leader when using temporary hold? [a]	Yes	32	89	13	76	10	100	9	100
	No	0	0	0	0	0	0	0	0
Annotated feedback									
Frequency during scoring	Frequently	5	14	4	24	1	10	0	0
	Occasionally	28	78	12	71	8	80	8	89
	Never	3	8	1	6	1	10	1	11
If never, would it be helpful?	Yes	4	100	1	100	2	100	1	100
	No	0	0	0	0	0	0	0	0
Was the frequency appropriate?	Too little	5	14	4	24	1	10	0	0
	Too much	1	3	0	0	0	0	1	11
	Just right	26	72	10	59	8	80	8	89
Did receiving annotated feedback affect scoring accuracy?	Yes	30	83	13	76	8	80	9	100
	No	4	11	3	18	1	10	0	0
Did receiving annotated feedback affect scoring rate?	Yes	19	53	7	41	5	50	7	78
	No	15	42	9	53	4	40	2	22

Note. Totals may not add to 100% due to nonresponses or rounding. ELPAC = English Language Proficiency Assessments for California.

[a] Temporary hold allows raters to request that a scoring leader review a response to determine what the assigned score should be.

All raters indicated that they used the chat feature during scoring, either frequently (31%) or occasionally (69%). A similar pattern was seen across all three programs. Just over one half (53%) of the raters indicated that the chat was initiated by both themselves and scoring leaders, with the remainder split between rater-initiated and scoring leader-initiated chats. Again, this pattern did not vary across different programs.

Most raters (69%) felt that using chat affected their future scoring accuracy. This was especially pronounced for ELPAC (76%) and GRE (80%) raters. Most raters commented that chat improved scoring accuracy because it provides more detailed and constructive feedback (e.g., "I'm getting more feedback now than I would with phone calls because it's short and quick and immediate. . . . This has increased my accuracy because I'm getting more feedback on responses"; "chat improves accuracy because I can have a conversation about a response I'm having trouble with. I can use information discussed in the future"), and as a result, raters were able to "self-correct" as needed.

However, the impact of chat on scoring rate was mixed. Most raters (58%) felt that chat affected their scoring rate. In particular, ELPAC (65%) and TOEFL (67%) raters felt that chat affected how fast they scored, while the majority of GRE raters (60%) felt it did not impact their scoring rate. This may point to differences in the nature of the responses being scored or to rater perceptions about scoring rate. That is, GRE responses are only in the form of essays, and GRE raters may feel that they are already scoring the essays at an optimal level. Some raters indicated that chat might distract them and cause them to slow down (e.g., "If you want to increase your accuracy you're going to slow down your rate"; "It does slow down scoring rate to be able to respond quickly. There was a constant back and forth that slows the process down"). Other raters felt that chat would be "a faster way to communicate with my SL," and some raters indicated that chat would have little impact because the time required to use it was minimal.

Most raters had no suggestions for improvements to chat, but some who did suggested correcting technical problems associated with chat, such as refreshing/clearing the message button of already-read messages and fixing the notification feature to make it obvious or more noticeable. Some raters also suggested adding a list of frequently asked questions for reference and adding a signal if a response from a scoring leader will be delayed.

Finally, all raters (89%) who used a temporary hold (where the rater moves his or her score to a response into a holding area to receive input and feedback from a scoring leader) indicated that they had communicated with the scoring leader when using a temporary hold. Generally, raters indicated that they put their comments into temporary hold and then used chat to let their scoring leader know they had put something on hold.

Scoring Leader Feedback

Raters next responded to a series of questions on feedback and other information they may have received from the scoring leader while performing operational scoring. These questions focused on the frequency, type, helpfulness, and consistency of the feedback. Table 3 presents the number and percentage of raters responding to each survey question option.

Table 3. Number and Percentage of Raters Responding to Each Survey Question Option on Scoring Leader Feedback While Scoring

Survey question	Response option	All raters		ELPAC raters		GRE raters		TOEFL raters	
		No.	%	No.	%	No.	%	No.	%
Frequency during scoring	Frequently	5	14	3	18	1	10	1	11
	Occasionally	22	61	12	71	3	30	7	78
	Rarely	8	22	2	12	5	50	1	11
How feedback was received									
E-mail	Yes	15	42	6	35	2	20	7	78
	No	20	56	11	65	7	70	2	22
Phone call	Yes	25	69	13	76	9	90	3	33
	No	11	31	4	24	1	10	6	67
IM	Yes	1	3	1	6	0	0	0	0
	No	35	97	16	94	10	100	9	100
Chat	Yes	35	97	16	94	10	100	9	100
	No	1	3	1	6	0	0	0	0
Phone text	Yes	0	0	0	0	0	0	0	0
	No	36	100	17	100	10	100	9	100
How helpful was scoring leader feedback?	Very helpful	25	69	11	65	8	80	6	67
	Somewhat helpful	10	28	5	29	2	20	3	33
	Not helpful	1	3	1	6	0	0	0	0
How consistent was scoring leader feedback?	Very consistent	14	39	5	29	3	30	6	67
	Somewhat consistent	18	50	9	53	6	60	3	33
	Not consistent	4	11	3	18	1	10	0	0

Note. Totals may not add to 100% due to nonresponses or rounding. ELPAC = English Language Proficiency Assessments for California; IM = instant message.

While the majority (61%) of raters indicated that they received feedback from a scoring leader occasionally while scoring, nearly one fourth of raters (22%) indicated that they rarely received feedback. In addition, while the majority of ELPAC raters (71%) and TOEFL raters (78%) reported that they received feedback occasionally, only about one third of the GRE raters (30%) reported occasional feedback, and half of GRE raters (50%) indicated that they rarely received feedback.

The most common method used by scoring leaders to communicate with raters was chat, as indicated by 97% of raters. Only a single ELPAC rater indicated not receiving feedback via chat. The second most common method was receiving a phone call, with 69% of raters indicating that they had received feedback in this way. In particular, the majority of GRE raters (90%) and ELPAC raters (76%) indicated that they had been called by a scoring leader. However, most TOEFL raters (67%) said that their scoring leader did not provide information via a phone call. The use of e-mail to provide feedback was split, with 42% of raters indicating that they had received feedback via e-mail and 56% indicating that they had not. TOEFL raters were the exception, with 78% indicating that e-mail was used by a scoring leader. Very few raters indicated that the other two methods—phone text and instant messaging (IM)—were used by a scoring leader. Only a very few received information via IM, and none of the raters received information via a phone text.

In general, scoring leader feedback was seen as very helpful (69%) or helpful (28%) by most raters. Only one ELPAC rater indicated that scoring leader feedback was not helpful. A greater percentage of GRE raters (80%) felt that scoring leader feedback was very helpful compared to ELPAC (65%) and TOEFL (67%) raters. Information provided by a scoring leader tended to relate to the score assigned by a rater (e.g., agreement with the assigned score, if score was too high or too low, reinforce scoring rubrics) or logistics related to operational scoring (e.g., a reminder to take a break, change in prompts).

One half of the raters (50%) felt that scoring leader feedback was somewhat consistent, although most TOEFL raters (67%) felt it was very consistent. Inconsistencies experienced by raters included the frequency of feedback (e.g., "The frequency to which they provide feedback varies somewhat"; "Some provide more feedback than others. Some will send out chat messages frequently and others not as often"), the level of detail ("Some provide more details, others just tell you 'this should be a score of 1' and that's it"), or content (e.g., "One will say the score is acceptable, then another says it is not—not for the same answer, but for similar answers").

Rater-Specific Feedback

The final set of questions focused on providing raters with information specific to their own performance as a rater. These questions focused on both individualized feedback and comparisons to other raters. Table 4 presents the number and percentage of raters responding to each survey question option.

Table 4. Number and Percentage of Raters Responding to Each Survey Question Option on Feedback on Rater's Own Performance

Survey question	Response option	All raters		ELPAC raters		GRE raters		TOEFL raters	
		No.	%	No.	%	No.	%	No.	%
Helpfulness of scoring accuracy	Very helpful	27	75	15	88	9	90	3	33
	Somewhat helpful	6	17	1	6	0	0	5	66
	Not helpful	3	8	1	6	1	10	1	11
Impact on future scoring accuracy	Yes	33	92	17	100	9	90	7	78
	No	2	6	0	0	1	10	1	11
Frequency for information	Hourly	17	47	10	59	3	30	4	44
	At beginning of scoring day	0	0	0	0	0	0	0	0
	At end of scoring day	6	17	2	12	2	20	2	22
	Other	13	36	5	29	5	50	3	33
Helpfulness of scoring rate	Very helpful	16	44	10	59	4	40	2	22
	Somewhat helpful	15	42	5	29	4	40	6	67
	Not helpful	5	14	2	12	2	20	1	11
Impact on future scoring rate	Yes	26	72	12	71	8	80	6	67
	No	10	28	5	29	2	20	3	33
Frequency for information	Hourly	13	36	8	47	2	20	3	33
	At beginning of scoring day	0	0	0	0	0	0	0	0
	At end of scoring day	8	22	3	18	3	30	2	22
	Other	14	39	6	35	5	50	3	33
Helpfulness of comparison to other raters	Very helpful	7	19	3	18	3	30	1	11
	Somewhat helpful	22	61	12	71	3	30	7	78
	Not helpful	7	19	2	12	4	40	1	11
Impact on future scoring accuracy	Yes	25	69	10	59	8	80	7	78
	No	10	28	6	35	2	20	2	22
Impact on future scoring rate	Yes	26	72	13	76	7	70	6	67
	No	9	25	3	18	3	30	3	33
How to provide information (Selected as first choice)	E-mail	1	3	0	0	0	0	1	11
	Link in ONE	8	22	4	24	4	40	0	0
	Dashboard in ONE	19	53	7	41	6	60	6	67
	Chat	7	19	6	35	0	0	1	11
	Phone call	1	3	0	0	0	0	1	11

Note. Totals may not add to 100% due to nonresponses or rounding. ELPAC = English Language Proficiency Assessments for California. ONE = Online Network for Evaluation.

The majority of raters (75%) said that knowing their own accuracy level while they were performing operational scoring would be very helpful; this was especially true for ELPAC raters (88%) and GRE raters (90%). However, only about one third of TOEFL raters (33%) indicated that having accuracy information would be very helpful, and the majority (66%) felt that it would be somewhat helpful. Overall, only a few raters (8%) felt that such information would not be helpful.

Generally, all raters (92%) felt that knowing their scoring accuracy would impact their future scoring accuracy. This was true for all three programs: ELPAC (100%), GRE (90%), and TOEFL (78%). Some raters felt that it would have a positive impact (e.g., "It would help give me a baseline for how I'm doing and help me know what I need to do to improve"; "You would know how to grade going forward. You'd feel more confident"; "I can make sure I do better next time"). However, not all raters indicated that knowing their scoring accuracy would have a positive impact on their future accuracy or were even sure that it would have an impact (e.g., "Minimally—for the most part it would hurt in a negative way. . . . You become a nervous wreck and [it] affects scoring"; "No—I don't see the correlation between those two things").

The ideal frequency for receiving such information varied quite a bit from program to program. More than half of ELPAC raters (59%) said they would want to know their accuracy levels hourly. For GRE, only about one third of raters (30%) indicated that they would want the information hourly, and for TOEFL, fewer than one half (44%). More than one fourth of ELPAC raters (29%), one half of GRE raters (50%), and one third of TOEFL raters (33%) indicated that they would want to know their accuracy rate more than once during a scoring day (e.g., two, three, or four times during the scoring day). Fewer raters indicated that they would want information on their accuracy level at the end of the scoring day (17% across all raters), and none indicated that they would want this information at the beginning of a scoring day.

Across all programs, fewer than one half (44%) of raters said that knowing their own scoring rate would be very helpful. While the majority of ELPAC raters (59%) felt that knowing their scoring rate would be very helpful, only 40% of GRE raters and 22% of TOEFL raters felt that way. TOEFL raters (67%) indicated that feedback on their scoring rate would be somewhat helpful, as did 40% of GRE raters.

In general, the majority of raters (72%) indicated that knowing their scoring rate would impact their future scoring rate. This finding was true across all programs, with 71% of ELPAC raters, 80% of GRE raters, and 67% of TOEFL raters saying that knowledge of their scoring rate would impact the speed at which they scored in the future. However, for many raters, how it would affect their scoring rate was not clear ("I think it could be good in some ways and bad in others. Might make you feel like you had to rush"; "Probably not because I try to maintain a constant rate anyway"; and "Unsure—there is a natural rhythm to scoring"). Some raters were confused about what the metric would represent ("My speed will vary on the depth and length of

the response, so that's sort of an arbitrary measure"; "It all depends how accurate the feedback is").

As with receiving information related to scoring accuracy, the ideal frequency for receiving scoring rate information varied from program to program. ELPAC raters (47%) tended to want scoring rate information hourly, while only 20% of GRE and 33% of TOEFL raters wanted hourly information. Some GRE (30%), TOEFL (22%), and ELPAC (18%) raters would want scoring information at the end of a scoring day; none indicated that they would want this information at the beginning of a scoring day. Other suggestions included two to three times during a shift, once at the beginning and again at the end of a shift, in 30 min intervals, and in the middle of a scoring shift.

When asked about the helpfulness of knowing how their performance, both accuracy and scoring rate, compared to other raters, the majority (61%) indicated that it would be somewhat helpful. This was especially true for ELPAC (71%) and TOEFL (78%) raters but not for GRE raters (30%). More GRE raters (40%) felt that such information would not be helpful.

The majority of raters (69%) felt that knowing how their accuracy compared to other raters would impact their future scoring accuracy. Some differences existed across programs, with 80% of GRE and 78% of TOEFL raters, but only 59% of ELPAC raters, indicating that it would impact their scoring accuracy. Raters felt it would affect them positively ("Because you are able to see where you are and others as well. We can make more accurate decisions"), and it would allow them to "have a better sense of what you're doing and know if you have to work harder." Others, however, felt that it would have a negative impact because it "would create competition and I hate competition. That would shut me down," and could become a distraction in that "[my] attention would not be 100% focused. Part of focus would be distracted by element of competition" or could result in increased anxiety (e.g., "I would get more anxious and that might mean I might overcompensate").

Some raters indicated that this would not impact their scoring accuracy because they were already scoring as accurately as they could (e.g., "If other people are more accurate than me, I don't know how it would make me more accurate"; "I strive to always maintain my accuracy and I take every response seriously. Knowing how other people are scoring is not going to affect how I approach that").

Raters also felt that knowing the rate at which other raters are scoring would impact their future scoring rate, with 75% of ELPAC, 70% of GRE, and 67% of TOEFL raters indicating this. Some felt it might result in a positive change (e.g., "If I'm noticing that my rate is either slower or faster than other raters, it might make me evaluate if I'm spending too much time on responses or not enough"; "It would let me know that I need to pick up the pace, and [give me] the motivation to do so"), while others felt it could be a negative influence (e.g., "Trying to speed up might affect accuracy"; "It would create a sense of competition to outpace those I am working with. This would make me anxious and probably lose it and maybe quit for the day"; "You really start to overanalyze and that could slow you down and affect your scoring in a negative way").

The majority of raters (53%) indicated that their first choice for receiving information about their scoring accuracy and scoring rate would be via a dashboard on the screen that automatically displays the information during scoring. Most GRE (60%) and TOEFL (67%) raters indicated this method to be their first choice, while only 41% of ELPAC raters indicated this preference. Receiving accuracy and scoring rate information via a link in ONE was the first choice for 40% of GRE raters, but only 24% of ELPAC raters and no TOEFL raters chose this method. Just over one third (35%) of ELPAC raters chose chat as the preferred method for receiving the information, while no GRE and only one TOEFL rater indicated this preference. Only two raters indicated either e-mail or a phone call as their first choice; both raters scored for TOEFL.

Summary of Results

Cognitive and learning sciences theory suggests that receiving the appropriate amount and type of feedback would impact raters' behavior, assist them with assigning accurate scores, and help maintain an acceptable level of performance while engaging in operational scoring. This study examined the perceived effectiveness of feedback practices used during rater training and operational scoring. While there is much consistency in how raters perceive these practices, some differences were apparent.

First, it should be recognized that there are some basic differences across the groups of raters and the tests they score. ELPAC raters score tests that are administered during specific time periods and generally are also California educators. The ELPAC test is given to K–12 students in California whose primary language is not English and is a fairly new test. The GRE and TOEFL assessments are established tests that are administered on a continuous basis to older

test takers. TOEFL test takers generally have English as a secondary language. The majority of GRE test takers' primary language is English, although many GRE test takers indicate English as a secondary language. In many cases, GRE and TOEFL raters are very experienced, having worked as raters for ETS over 5 or more years. Thus, raters' perceptions about the usefulness of particular types of feedback may be the result of the nature of the test, characteristics of the test takers, and their own scoring or occupational experience.

The majority of raters indicated that they did not use the chat feature during training, with a number of raters indicating that they were trained prior to chat being available. However, these raters still felt that having the opportunity to use it during training would have been helpful.

However, during scoring, all raters indicated that they used chat. Most raters (89%) viewed chat in a positive light, and many commented that using chat allowed them to get immediate feedback from a scoring leader. Raters also felt that chat improved their scoring accuracy because it provided more detailed and constructive feedback, and as a result, raters' self-confidence increased, and they were able to "self-correct" as needed. A few raters cautioned that chat could have a negative impact on scoring accuracy if the response took too much time to get to the rater. A few negative comments about chat were also provided by raters, such as it being distracting, the need to type everything as part of the chat, the possibility for misunderstanding the text in the chat, and the chat feature not being in the same window as the response being scored.

While chat was seen as positive during training and operational scoring, a number of raters felt that there were risks involved with introducing the chat feature as part of calibration. In particular, many felt that chat would detract from the purpose of calibration; that is, calibration was meant to be a quick test that measured how much raters learned as part of training. In addition, some raters felt the use of chat could delay qualified raters moving on to operational scoring.

Overall, most raters received feedback directly from a scoring leader while performing scoring on at least an occasional basis. However, one half of GRE raters indicated that they rarely received feedback. This might reflect the relative experience level of GRE raters or program-specific administrative requirements (e.g., ratio of scoring leader to raters; number of back-readings done by scoring leaders). All but one rater indicated that scoring leader feedback was somewhat or very helpful. However, the majority of raters felt that the feedback was only

somewhat or not consistent. Feedback received via chat appeared to be the most consistent, possibly because raters could ask for clarification of the feedback as part of chat. While chat was the more frequent method for receiving feedback, a substantial number of raters also received feedback via a phone call or e-mail from the scoring leader.

The type of feedback received also was perceived as important by raters. During training and operational scoring, most raters had received annotated feedback and felt that additional feedback of this type would be helpful. The majority of raters indicated that their accuracy level would likely increase if feedback included an explanation of why their score was incorrect.

Raters expressed mixed reactions to receiving information about their own scoring rate, their accuracy level, or how they compared with other raters. The majority of raters (99%) indicated that knowing their own level of accuracy would likely have some positive impact on performance—depending on the type of information that was given. Raters felt they would be more likely to self-correct and increase their self-confidence level if detailed and constructive feedback (such as annotated feedback) was given and not just a single number that represented accuracy level. Some raters felt that this type of feedback would also be used as a reference while scoring in the future.

However, some raters (28%) felt that knowing their accuracy level would have a negative impact on performance. They indicated that their anxiety level would increase and confidence level decrease, which would further impact accuracy levels. One rater felt that receiving information about scoring accuracy while performing scoring would result in raters paying more attention to the number of responses they were scoring, not accuracy.

The majority of raters (89%) also felt that knowing their scoring rate would influence their future scoring rate. However, it was not clear whether knowing their scoring rate would have a positive or negative influence on how raters would perform in the future. Comments from raters indicated concerns ("it would make you rush") as well as confusion about the metric ("an arbitrary measure").

Knowing how a rater's scoring accuracy level compared with other raters' accuracies received mixed reactions. While the majority of raters (89%) indicated that having a comparison to other raters would be helpful, some indicated it could have a positive impact on their scoring accuracy, while others indicated it could have a negative impact (e.g., it would create competition among raters, be distracting, or cause anxiety).

Raters also felt that knowing how fast other raters were scoring would impact their own scoring rate. Again, most raters felt that such information could be positive ("pick up my pace"), but some raters felt that such information could be negative ("create competition" and "affect accuracy").

Differences in the frequency of feedback and the method for providing the feedback may be dependent on the experience level of the rater. The ideal frequency for receiving information on scoring accuracy level and scoring rate varied depending upon the program for which the rater scored. Raters with less experience, as reflected in responses from ELPAC raters, appear to desire feedback more frequently than those with more experience (GRE and TOEFL raters). ELPAC raters indicated that they would want such information more frequently (hourly, two to three times a scoring shift) than did GRE or TOEFL raters. Again, this may reflect the characteristics of the raters or the test takers, or the stakes associated with the use of the scores. However, none of the raters indicated that they would want this information at the beginning of their scoring shift; such information seems to be more useful once raters are actively scoring.

About one half of raters indicated that they would want information about scoring rate and scoring accuracy to be given via a dashboard on the screen that automatically displays the information while they are scoring. However, fewer ELPAC raters wanted the information via dashboard compared to GRE and TOEFL raters. ELPAC raters were almost equally divided between receiving the information via dashboard and receiving the information via chat. GRE raters indicated receiving the information via a dashboard or via a link in ONE that took them to the information to be their preferred choice. For TOEFL raters, the dashboard was overwhelmingly their preferred choice. Thus the method of providing the feedback must be easily accessible—either displayed on the screen while scoring or easily obtained through a link.

Conclusion

The results of this study indicate that the level, type, and frequency of feedback define its usefulness to raters. Raters indicated that feedback given during training or operational scoring is valuable, but raters also indicated that to be valuable, the feedback needs to be immediate and concise. For example, raters felt that the most helpful aspect of chat during training and operational scoring was that chat provides immediate feedback and allows for specific explanations and discussion. Overall, feedback from scoring leaders was perceived as valuable, regardless of how it was provided to raters (e.g., chat, e-mail, phone call).

However, providing the same type of feedback that is given during training or scoring as part of calibration was perceived as being undesirable. Many raters indicated that providing any feedback during calibration would circumvent the purpose of calibration.

To be useful, feedback designed to positively impact scoring accuracy needs to be swift and specific and to provide specific information that indicates why a rater's assigned score is incorrect. Many raters felt that receiving feedback on their own scoring accuracy level would serve to make them personally accountable for improvement, but only if the feedback was detailed and specific, and some raters cautioned that such information could be a detriment to scoring accuracy levels in that raters could become more anxious and less confident as they score.

In addition, receiving information on their own scoring rate was perceived as positive by raters because it would allow them to self-regulate their speed (i.e., to slow down or speed up). Some raters, however, indicated that it would have no impact because raters can only score at a certain pace and remain accurate. Finally, feedback on scoring rate needs to be more than a number—it needs to provide information that is provided in a context that it is easily interpretable and understandable by raters.

Some raters expressed caution about providing information about scoring accuracy levels and scoring rate during operational scoring, especially when the rater is compared to other raters, indicating that such information could make raters competitive or anxious, which could result in less accurate scores. In addition, there were perceived differences in how the same type of feedback might impact scoring accuracy level and scoring rate. For example, raters felt that providing annotated feedback during operational scoring could increase scoring accuracy levels, but raters also indicated that providing this type of feedback could reduce read rates.

In summary, feedback must be timely—but not necessarily continually provided—and give appropriate information that allows raters to understand what they did incorrectly so that they can self-correct. However, feedback seen as valuable during training and operational scoring is not desirable during calibration because the role of calibration is to determine if raters have acquired the appropriate level of knowledge to score. Caution must also be taken that providing feedback does not set up a "competition" among raters or that it creates anxiety and doubt among raters who are performing well. Finally, how to best explain what performance

measures, such as accuracy level and read rate, represent and how raters should interpret such measures must be carefully considered to increase and maintain rater performance.

References

Arthur, W., Jr., Bennett, W., Jr., Stanush, P. L., & McNelly, T. L. (1998). Factors that influence skill decay and retention: A quantitative review and analysis. *Human Performance, 11*, 57–101. https://doi.org/10.1207/s15327043hup1101_3

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Bejar, I. I. (2017). A historical survey of ETS research regarding constructed-response formats. In R. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 565–633). New York, NY: Springer. https://doi.org/10.1007/978-3-319-58689-2_18

Braun, H. I. (1988). Understanding score reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*, 1–18. https://doi.org/10.3102/10769986013001001

Cason, G. J., & Cason, C. L. (1984). A deterministic theory of clinical performance rating. *Evaluation and the Health Professions, 7*, 221–247. https://doi.org/10.1177/016327878400700207

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Erlbaum.

Finn, B., Thomas, R., & Rawson, K. A. (2018). Learning more from feedback: Elaborating feedback with examples enhances concept learning. *Learning and Instruction, 54*, 104–113. https://doi.org/10.1016/j.learninstruc.2017.08.007

Finn, B., Wendler, C., & Arslan, B. (2018, April). *Applying cognitive theory to the human essay rating process.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Finn, B., Wendler, C., Ricker-Pedley, K., & Arslan, B. (2018). *Does the time between scoring sessions impact scoring accuracy? An evaluation of constructed-response essay responses on the GRE General Test* (Research Report No. RR-18-31). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12217

Latham, G. P., & Locke, E. A. (2007). New developments in and directions for goal-setting research. *European Psychologist, 12*, 290–300. https://doi.org/10.1027/1016-9040.12.4.290

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice Hall.

Morisano, D., Hirsh, J. B., Peterson, J. B., Pihl, R. O., & Shore, B. M. (2010). Setting, elaborating, and reflecting on personal goals improves academic performance. *Journal of Applied Psychology, 95*, 255–264. https://doi.org/10.1037/a0018478

Pashler, H., Cepeda, N., Wixted, J., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 3–8. https://doi.org/10.1037/0278-7393.31.1.3

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

Skinner, B. F. (1953). *Science and human behavior*. New York, NY: Simon and Schuster.

Wendler, C., Glazer, N., & Bridgeman, B. (2018, April). *The impact of setting scoring expectations on scoring rates and accuracy.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Wiese, B. S., & Freund, A. M. (2005). Goal progress makes one happy, or does it? Longitudinal findings from the work domain. *Journal of Occupational and Organizational Psychology, 78*, 287–304. https://doi.org/10.1348/096317905X26714

Wolfe, E. W. (2014). *Methods for monitoring rating quality: Current practices and suggested changes*. Retrieved from the Pearson website: https://www.pearson.com/corporate/efficacy-and-research/schools-education-research/research-reports/assessment/issues-in-assessment.html

Wolfe, E. W., Winchester, E. P., & Rupp, A. A. (2018). *The effectiveness of formative feedback for essay raters* (CR Scoring Research Brief No. 2018-02, Statistical Report No. SR-2018-029). Princeton, NJ: Educational Testing Service.

Appendix: Rater Feedback Survey

1. Do you score for ETS only or for ETS and another company?
2. What programs have you scored for ETS?
3. As you went through RATER TRAINING, how often did you use the chat feature? Did you use it frequently, occasionally, or never?

   ____ Frequently
   ____ Occasionally
   ____ Never

   Is there a particular reason you did not use it during training?
   Do you believe using chat during training would be helpful?
   If no: Why do you believe it would not be helpful?
   If yes: Why do you believe it would be helpful?
4. What did you find to be the MOST helpful about the chat feature during training?
5. What did you find to be the LEAST helpful about the chat feature during training?
6. How could ETS improve the chat feature?
7. After you completed your training, you were required to pass CALIBRATION. Currently the ability to chat with your scoring leader is not possible during calibration. Do you think having the chat feature during calibration would be helpful?

   If no: Why do you believe it would not be helpful?
   If yes: Why do you believe it would be helpful?
8. As you went through RATER TRAINING, how often did you receive annotated feedback? Did you receive it frequently, occasionally, or never?

   ____ Frequently
   ____ Occasionally
   ____ Never

   Do you believe annotated feedback would have been helpful?
   If no: Why do you believe it would not have been helpful?

If yes: Why do you think it would have been helpful?

9. During training, was the frequency of annotated feedback too little, too much, or just right?

   ____ Too little
   ____ Too much
   ____ Just right

10. Currently annotated feedback is not provided during CALIBRATION. Do you think receiving annotated feedback during calibration would be helpful?

    If no: Why do you believe it would not be helpful?
    If yes: Why do you believe it would be helpful?

11. How often did you use the chat feature while SCORING? Did you use it frequently, occasionally, or never?

    ____ Frequently
    ____ Occasionally
    ____ Never

    Is there a particular reason you did not use it while scoring?

12. Who initiated the chat?

    ____ Rater
    ____ Scoring leader
    ____ Both

    Can you give me an example of when YOU initiated the chat?
    Can you give me an example of when your SCORING LEADER initiated the chat?

13. What did you find to be the MOST helpful about the chat feature while you were scoring?

14. What did you find to be the LEAST helpful about the chat feature while you were scoring?

15. Do you believe the use of the chat feature affected your future level of scoring accuracy?

    If yes: How so?

    If no: Why do you believe it didn't?

16. Do you believe the use of the chat feature affected your scoring rate?

    If yes: How so?

    If no: Why do you believe it didn't?

17. How could ETS improve the chat feature during scoring?

18. While you were scoring, did you communicate with your scoring leader in any way when you put a score into TEMPORARY HOLD?

    ___ Never used temp hold

    ___ No

    ___ Yes

    Can you explain how you communicated with your scoring leader?

19. How often did you receive annotated feedback while you were scoring? Would you say frequently, occasionally, or never?

    ____ Frequently

    ____ Occasionally

    ____ Never

    Do you believe receiving annotated feedback would have been helpful?

    If no: Why do you believe it would not have been helpful?

    If yes: Why do you believe it would have been helpful?

20. While scoring, was the frequency of annotated feedback too little, too much, or just right?

    ____ Too little

    ____ Too much

    ____ Just right

21. Do you believe that receiving annotated feedback affected your future scoring accuracy?

If yes: How so?

If no: Why do you believe it didn't?

22. Do you believe that receiving annotated feedback affected your scoring rate?

    If yes: How so?

    If no: Why do you believe it didn't?

23. During a typical scoring session, how frequently do you receive feedback from your scoring leader? Do you feel you receive feedback frequently, occasionally, or rarely?

    ____ Frequently
    ____ Occasionally
    ____ Rarely

24. What type of information do you generally receive from your scoring leader?

25. Have you received feedback from your Scoring Leader via:

	Yes	No
E-mail?	____	____
Phone call?	____	____
IM?	____	____
Chat?	____	____
Phone text?	____	____
Any other?		

26. In general, how helpful have you found the feedback from your scoring leader to be? Have you found it very helpful, somewhat helpful, or not helpful?

    ____ Very helpful
    ____ Somewhat helpful
    ____ Not helpful

27. You may have had a different scoring leader during different days you score. How consistent have you found the feedback from different scoring leaders to be? Have you found it to be very consistent, somewhat consistent, or not consistent?

    ____ Always have had the same scoring leader each day

____ Very consistent

____ Somewhat consistent

What kinds of things were inconsistent?

____ Not consistent

What kinds of things were inconsistent?

28. During scoring, would it be helpful to see how accurately you are scoring? Would it be very helpful, somewhat helpful, or not helpful?

    ____ Very helpful

    ____ Somewhat helpful

    ____ Not helpful

29. Do you believe knowing how accurately you are scoring would affect your future level of scoring accuracy?

    If yes: Why do you think it would affect how accurately you subsequently score?

    If no: Why do you believe it wouldn't?

30. How frequently would you want this information during scoring: hourly, once at the beginning of the scoring day, once at the end of the scoring day, or some other?

    ____ Hourly

    ____ At beginning

    ____ At end

    ____ Other

    Could you describe this?

31. During scoring, how helpful would it be to know your scoring rate? Would it be very helpful, somewhat helpful, or not helpful?

    ____ Very helpful

    ____ Somewhat helpful

    ____ Not helpful

32. Do you believe knowing your scoring rate would affect the rate at which you score?

    If yes: Why is that?

If no: Why do you believe it wouldn't?

33. How frequently would you want this information during scoring: hourly, once at the beginning of the scoring day, once at the end of the scoring day, or some other?

    ____ Hourly

    ____ At beginning

    ____ At end

    ____ Other

    Could you describe this?

34. During scoring, how helpful would it be to know how your performance—accuracy and scoring rate—COMPARED TO OTHER RATERS? Would it be very helpful, somewhat helpful, or not helpful?

    ____ Very helpful

    ____ Somewhat helpful

    ____ Not helpful

35. Do you believe this comparison would affect how ACCURATELY you score?

    If yes: Why is that?

    If no: Why do you believe it wouldn't?

36. Do you believe this comparison would affect your SCORING RATE?

    If yes: Why is that?

    If no: Why do you believe it wouldn't?

37. There could be different ways of providing you with information about your scoring accuracy and scoring rate. For example, there could be a link in ONE that would take you to the information, a dashboard on the screen that automatically displays this information as you are scoring, or the scoring leader could send you a separate e-mail, contact you via chat, or call you.

    Please list these in priority order, with 1 meaning the MOST desirable and 5 meaning the LEAST desirable: e-mail, link, dashboard, chat, or phone call.

    ____ E-mail

    ____ Link

____ Dashboard

____ Chat

____ Phone call

38. Finally, is there other information you feel would be helpful to know as you score?