ETS GRE

Assess Ability. Predict Performan

ETS GRE® Board Research Report

ETS GRE®–19-01

ETS RR–19-09

Examining the Calibration Process for Raters of the *GRE*® General Test

Cathy Wendler

Nancy Glazer

Frederick Cline

December 2019

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

****

****

GRE-ETS

PO Box 6000

Princeton, NJ 08541-6000

USA

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

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

RESEARCH REPORT

Examining the Calibration Process for Raters of the *GRE*® General Test

Cathy Wendler, Nancy Glazer, & Frederick Cline

Educational Testing Service, Princeton, NJ

One of the challenges in scoring constructed-response (CR) items and tasks is ensuring that rater drift does not occur during or across scoring windows. Rater drift reflects changes in how raters interpret and use established scoring criteria to assign essay scores. Calibration is a process used to help control rater drift and, as such, serves as a type of quality control during CR scoring. Calibration sets are designed to provide sufficient evidence that raters have understood and internalized the rubrics and can score accurately across all score points of the score scale. This study examined the calibration process used to qualify raters to score essays from the *GRE*® Analytical Writing measure. A total of 46 experienced raters participated in the study, and each rater scored up to 630 essays from 1 of 2 essay prompt types. Two research questions were evaluated: *Does calibration influence scoring accuracy?* and *Does reducing the frequency of calibration impact scoring accuracy?* While the distribution of score points represented by the essays used in the study did not necessarily reflect what raters see during operational scoring, results suggest that the influence of calibration on Day 1 remains with raters through at least 3 scoring days. Results further suggest that scoring accuracy may be moderated by prompt type. Nevertheless, study results indicate that daily calibration for GRE prompt types may not be necessary and that reducing the frequency of calibration is unlikely to reduce scoring accuracy.

Keywords *GRE*® general test; calibration; CR scoring; human raters

doi:10.1002/ets2.12245

The use of constructed-response (CR) tasks, such as essays, speech samples, and short answers, which require test takers to construct a response rather than choose an answer from a predetermined set of options, continues to grow, and new task and item types utilizing human raters and/or automated scoring engines continue to be developed. However, while some CR tasks and item types lend themselves to automated scoring, human raters are likely to remain an integral part of CR scoring in the near future. Challenges associated with scoring CRs are not new (see Bejar, 2017), and one of the biggest challenges in scoring CRs that use human raters continues to be 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.

Human ratings are subject to rater variability that may impact the reliability of scores (Braun, 1988). The variability associated with human raters, generally termed *rater effects*, refers to scoring patterns that reflect measurement error related to raters (Engelhard, 2002; Wolfe, 2014). Zhang (2013) described some common rater errors and biases in CR scoring, such as scoring too severely or too leniently compared to other raters (*severity effect*), scoring test takers' responses too harshly or leniently on the basis of other variables not necessarily related to the construct being scored (*bias effect*), failing to use the extreme ends of the score scale (*central tendency effect*), and applying scoring criteria inconsistently over time (*rater drift*). Past research has shown that rater errors explain as much as or more score variability as test taker ability (Cason & Cason, 1984). Therefore controlling for and monitoring rater drift is critically important to ensure valid scores.

The use of prescored responses — referred to as *exemplar responses* — for qualifying and monitoring raters is common as part of the scoring process (Glazer, 2017). In the context of the current study, exemplar responses are essays written by test takers that have been selected, reviewed independently by experts, determined to be clear examples of various score levels that align with the established scoring rubrics, and assigned that particular score. Thus exemplar essays are considered the "gold standard" by which rater performance can be monitored (Ricker-Pedley, 2011).

Exemplar responses may be used in many ways. For example, they may be used as *benchmarks*, sometimes referred to as *anchors*, that demonstrate the characteristics of an essay at a particular score level (Parke, Lane, & Stone, 2006); assembled

Corresponding author: C. Wendler, E-mail: cwendler@ets.org

together to form *certification tests* following rater training to determine if a rater is ready to move into operational scoring; used as *validity* (or *monitoring*) *samples* that are seeded into the scoring system along with operational (i.e., real) responses of test takers and used as a mechanism for monitoring rater performance during operational scoring; and used in *calibration sets* given at the beginning of or during a scoring cycle to provide evidence that raters have understood and internalized the training responses and the rubric and can accurately categorize responses across all score points of the score scale.

Lane and Stone (2006) provided a generic description of CR scoring procedures. Generally, the process of ensuring and monitoring rater quality begins with rater training. Training frequently consists of self-study and practice on exemplar responses and provides raters with practice scoring a variety of responses that illustrate important features of the scoring rubric (Everson & Hines, 2010). Once raters complete training, they score exemplar responses in a preassembled calibration set, and if they successfully complete calibration, they may move on to score test takers' operational responses.

Rater calibration is a process used to help control *rater drift* (Congdon & McQueen, 2000; Lumley & McNamara, 1995). As such, the calibration process serves as a type of quality control during the CR scoring window (McClellan, 2010; Myford & Wolfe, 2009). Calibration sets are designed to provide sufficient evidence that raters have understood and internalized the rubrics and can score accurately across all score points in the score scale. Thus, calibration plays a dual role: It ensures that multiple examples of each score point are displayed for raters, thereby reinforcing the scoring criteria raters will need to know and apply, and it also serves a "gatekeeping" function, as raters who fail to achieve an expected level of performance will not be allowed to do operational scoring (Glazer, 2017).

Many testing programs require that raters pass calibration at the beginning of each scoring day or when the type of CR being scored changes. Raters are not allowed to perform operational scoring if they fail to meet certain performance standards as part of calibration. Criteria used to determine a "pass" on calibration vary across programs and tests but generally include the extent to which the rater's scores agree with exemplar responses' assigned scores. A specific level of performance (e.g., percentage of exact agreement with each exemplar response's score, number of allowable discrepant scores [i.e., scores that deviate by a certain number of score points from the exemplar response's assigned score]) generally constitutes the criteria needed for passing calibration.

However, time spent in calibration may be considered nonproductive because it does not result in usable, reportable scores. Raters who fail calibration are often still paid for their time, even though they are not allowed to score operational essays, resulting in additional costs for CR scoring. However, this logistical consideration needs to be weighed in terms of the benefits received from calibration: the more reliable the raters, the fairer the scores.

Some studies have shown that successfully completing calibration is predictive of accurate scoring on the same day that calibration occurs (Ricker-Pedley, 2011; Ricker-Pedley & Li, 2010). Other studies have suggested that calibration may be influenced by other variables, such as time between operational scoring.

For example, Finn, Wendler, and Arslan (2018) and Finn, Wendler, Ricker-Pedley, and Arslan (2018) examined a number of variables that may influence accurate scoring, such as the number of days in a scoring gap, for essay responses to both the *TOEFL*® test and the *GRE*® general test essay responses. Among other findings, results suggested that increasing the number of days in a scoring gap has a negative influence on calibration performance for both GRE and TOEFL and, for GRE, operational performance. They also found that raters tend to spend more time scoring essays in a calibration set than they do scoring essays that are part of operational scoring.

A study by Wendler and Cline (2017) examined rater performance using operational data across a 5-day scoring window for raters who calibrated only on their first day of scoring. Results indicated only a slight decrease in overall agreement rates from Day 1 to Day 2 for two of the CR tasks. For the third CR task, agreement rates remained steady or slightly increased in subsequent scoring days. Results also suggested that the ability of raters to reliably classify responses as belonging to a particular score level did not appear to vary less across days for raters who scored every day compared to those raters who had a scoring gap. Overall, agreement rates appeared to be more dependent on the particular CR task rather than on time from calibration.

These past studies showed the value of calibration in controlling rater drift (Congdon & McQueen, 2000; Lumley & McNamara, 1995; Ricker-Pedley, 2011; Ricker-Pedley & Li, 2010). However, for the most part, these studies focused on scoring accuracy on the same day that calibration occurred, whereas the current study also examined performance on subsequent scoring days.

Results of the Wendler and Cline (2017) study suggested that even when calibration only occurs on the first day of a multiday scoring event, agreement rates remain steady or even slightly increase on subsequent scoring days. However, this study was different from the current study in that it used operational data only and could not control for the type of prompt assigned to each rater. In addition, the composition of the group of raters differed each scoring day, and daily results therefore were based on average performance across a very large, but slightly different, group of raters. Finally, given that the test takers were younger than those who take the GRE, the essays evaluated in the study were less complex and shorter than those generally produced by GRE test takers.

The GRE Calibration Process

The GRE Analytical Writing measure assesses critical thinking and analytical writing skills. It tests test takers' ability to articulate and support complex ideas, construct and evaluate arguments, and create focused and coherent discussions. Test takers compose essays in response to two types of prompts: argument and issue. The argument prompt requires test takers to assess the claims made in an argument that is presented to them and to evaluate if the provided evidence supports the argument. In their essay, test takers consider the logical nature of the argument rather than agreeing or disagreeing with the position it presents. The issue prompt requires test takers to develop a response that provides evidence and examples that support their view on a general issue. The two task types are complementary in that one requires test takers to construct their own argument and the other requires them to evaluate someone else's argument.

Essay variants may be created from the same prompt. Essay variants are multiple versions of a prompt with similar or identical wording but with different instructions as to how test takers should respond to the prompt. Test takers must attend to the specific task directions and compose essays in response to the directions.

GRE raters are required to go through calibration at the beginning of their daily scoring session. This step is required for all raters, regardless of their scoring experience, amount of time since they last scored, or prior individual performance levels. Raters generally are scheduled to score essays from only the argument or issue prompt type during a single scoring session. However, raters may be asked, although infrequently, to switch from one prompt type to the other during a single scoring session. In that case, raters are required to calibrate again because the scoring rubrics are unique to the prompt type.

GRE raters are given two attempts to pass calibration. Currently, each calibration set consists of 10 exemplar essays. Raters are not given a time limit for completing calibration, although most raters take 20–40 minutes.

To pass calibration, GRE raters must classify at least 6 of the 10 exemplar essays into the same category as the exemplar essay's assigned score (i.e., at least 60% exact agreement) and are allowed no discrepancies (defined as differing from the exemplar essay's assigned score by 2 or more score points). The established classification rate of 60% exact agreement represents the minimal level that is believed to be needed to ensure that a rater is accurately applying the scoring rubrics. Raters who fail both attempts may not score that day but may attempt calibration again at their next scoring session (often the next day).

Both human raters and automated scoring (i.e., the *e-rater*® automated scoring engine) are used to operationally score GRE essays. Each essay receives a score from a human rater and an e-rater score. If the two scores do not agree within a specified threshold, the essay is scored by another human rater. This adjudication process helps ensure that all essay responses are scored accurately.

This study was designed to evaluate the impact of calibration frequency on rater drift, defined as changes in the level of exact agreement with the exemplar essay score. The study focused on two general questions:

- Does calibration influence scoring accuracy
    - on the same scoring day?
    - on subsequent scoring days?
    - similarly for both the argument and issue prompts?
- Does reducing the frequency of calibration impact scoring accuracy?
    - Does scoring accuracy decline with increasing time from calibration?
    - Are results similar for both the argument and issue prompts?

Study Design and Method

Study Design

The study involved scoring GRE essays over a predefined period; trained and experienced GRE raters were the participants in the study. Raters who had scored for GRE during previous operational scoring windows were invited to participate in the study. A total of 50 raters were selected from among the volunteers. Raters were randomly placed into one of two prompt-type groups for scoring: argument or issue.

The study took place over a 5-day period. Calibration was required on the first day only, and raters who did not pass calibration on the first day were dropped from the study. Raters who passed calibration on the first day were allowed to score for a total of 5 days without further calibration. A total of 46 raters, 23 in each prompt-type group, ultimately participated in the study.[1]

Raters were told that they were part of a GRE essay scoring project and that ETS Assessment Division staff, rather than scoring leaders, would be monitoring their scoring and progress. It was not made explicit whether the essays were operational. Raters were compensated at their usual rate of pay. After the 5 scoring days, raters were scheduled for operational scoring and were again required to pass daily calibration.

All scoring was completed using the ONE system, an ETS proprietary system that enables raters to score responses to many types of CR tasks via secure Internet access. Raters had access to the same scoring aids they have when performing operational scoring. For GRE, raters use a holistic approach to score each essay on a 6-point scale. Holistic scoring uses scoring rubrics to guide raters in making a single judgment of a response as a whole. The assigned score reflects the integration of a number of discrete features specified by the rubric (Baldwin, Fowles, & Livingston, 2008) and reflects the overall performance using the specific criteria (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 2014, p. 219) that underlie the scoring rubrics.

Two different prompts were used each day to help control for prompt-specific effects for a total of 10 argument and 10 issue prompts. All prompts came from the same variant group (i.e., Group 1). This variant group was determined by assessment experts to be the most generic and therefore the most likely to produce results that would generalize to the other two variant groups (Bridgeman, Trapani, & Bivens-Tatum, 2011).

Exemplar essays that had been used in calibration sets or as validity samples formed the pool of essays that were scored. Because these essays were already vetted through experts, it eliminated the need for staff to make judgments about the appropriate scores of the essays. Each day of the study, raters received the exemplar essays in the same order.

Raters scored for up to 6 hours each day.[2] Each day raters were allowed to score essays written in response to the first prompt for 3 hours and then were switched to the second prompt for the next 3 hours. Between 124 and 134 essays per prompt were provided each day for scoring. Figure 1 displays the study design.

The week following the study, raters were sent a short survey created for the study asking for their opinion about different aspects of the GRE scoring project (see Appendix A). No demographic information about the raters was collected; therefore, only limited comparisons across raters could be made. Completing the survey was voluntary, and raters were not compensated for their time. Nevertheless, 32 of the 46 raters responded to the survey. Finally, because the survey was anonymous, rater performance in the study could not be connected with their survey responses.

Rater Scoring Quality

Various statistics of rater scoring quality exist, such as the percentage of exact, adjacent, and discrepant agreement with the assigned score for each exemplar response (Williamson, Xi, & Breyer, 2012); agreement rates with other raters (kappa, weighted kappa); and more sophisticated measures, such as the point biserial correlation (Attali, 2016). For this study, we calculated several indices to investigate rater scoring quality: (a) the percentage of exact agreement with the assigned score for each exemplar essay, (b) Cohen's kappa, and (c) quadratic weighted kappa (QWK).

Although exact + adjacent (where *adjacent* means the rater is one below or above the exemplar essay score category) may also be used as a measure of rater performance, the exact + adjacent agreement rates were so high for these raters that agreement rates appeared to be nearly identical across days. For example, exact + adjacent agreement ranged from .998 to .994 for the argument prompt and from .990 to .996 for the issue prompt. As a result, the analyses reported on in this study use only exact agreement.

Prompt Type	Number of Raters	Day 1 Calibration	Day 2 No Calibration	Day 3 No Calibration	Day 4 No Calibration	Day 5 No Calibration
Argument	23	Prompt A-1 Prompt B-1	Prompt A-2 Prompt B-2	Prompt A-3 Prompt B-3	Prompt A-4 Prompt B-4	Prompt A-5 Prompt B-5
	Number of Available Essays	124	134	124	124	124
Issue	23	Prompt C-1 Prompt D-1	Prompt C-2 Prompt D-2	Prompt C-3 Prompt D-3	Prompt C-4 Prompt D-4	Prompt C-5 Prompt D-5
	Number of Available Essays	124	134	124	124	124

Figure 1 Study Design.

Kappa and QWK are common measures of interrater agreement for qualitative items and are chance-corrected agreement indices (i.e., both take into account the possibility of agreement occurring just by chance). QWK also takes into account nonexact scores and assigns a weight to the nonexact scores when computing rater agreement.

Scoring quality for each rater was determined by comparing the score each rater assigned to an essay with each exemplar essay's assigned score. Therefore, the assigned score for the exemplar essay functioned as a second rater's score. Agreement rates were calculated by determining, for each rater, the percentage of exact agreement with each exemplar essay's assigned score. The average percentage exact agreement rate was then calculated by day and prompt type. Thus exact agreement reflects how frequently raters classified the exemplar essay into the same score category that had been assigned to the essay. We use the term *agreement rate* to refer to the classification accuracy in the remainder of this study.

In addition, kappa and QWK were computed for each study day, again using each rater as Rater 1 and the exemplar essay score as Rater 2.

Most, but not all, raters completed scoring all essays within the allotted time frame. However, because raters score at different rates and because essays were only presented once on a specific day, not every rater was able to score all the essays each day. As a result, when examining exact agreement rates by day, more efficient (faster) raters contributed more to the overall agreement rate than less efficient (slower) raters. For this reason, the majority of the analyses use agreement rates at the essay level as the unit of analysis, not an aggregate of agreement rates by rater.

As agreement rates between raters' scores and the exemplar essays' scores were calculated, we found that, in some instances, there was a low level of exact agreement. There are several possible reasons for this, such as differences in prompt-level characteristics and differences in the number of hours raters were allowed to score responses to each prompt. Another concern was that the scale points represented by the essays used in the study did not accurately reflect what raters would expect to see operationally. That is, the number of essays representing each of the six score points on the scale was equal; in operational scoring, comparatively few essays are scored as 1 or 6. As a result, even accurate raters might, either consciously or unconsciously, adjust their scoring to reflect their expectations of how common scores of 1 and 6 have been in the past. For this reason, analyses were also conducted by score level.

One possible approach was to treat essays with low exact agreement rates in the study as "bad items" and to remove them from the analysis. However, because essays used as exemplar responses may be selected for their specific attributes, agreement rates may be affected by these attributes. In this case, rater agreement in the study may simply reflect the low level of agreement seen when these difficult-to-score essays were used operationally. When used operationally, these essays would have been spread out across a number of scoring windows. If, however, these difficult-to-score essays were, by chance, all used in the study on the same day, agreement rates for that day would be affected.

To account for the potential variation in the difficulty of scoring the exemplar essays, the exact agreement rates seen when the exemplar essays were used in operational scoring were examined. Essays used in the current study that had been scored by at least 20 raters during operational scoring were identified, resulting in about 42% ($n = 526$) of the essays meeting this criterion.[3] For these essays, an operational exact agreement rate was determined by calculating the percentage of exact agreement for each exemplar essay when used during operational scoring.

A new exact agreement rate—referred to here as *modified agreement*—was calculated by first dropping those essays that were not scored by at least 20 raters during operational scoring, recalculating the mean exact agreement using the remaining essays for each rater, and subtracting the operational exact agreement rate from the mean agreement rate for each study day. In effect, this modified agreement rate represented the overall difficulty of the task each day of the study, including differences in the numbers of essays scored by an individual rater.

While this allowed us to take into account the ease or difficulty of scoring each exemplar essay, it also meant that agreement rates were based on fewer than one half of the scored essays. For this reason, analyses were conducted separately, once using all available essays and once using only those essays where operational data were available.

Results

Agreement Rates

Table 1 displays the mean exact agreement rate by study day and prompt type using all exemplar essays scored during the study. Note that the number of scores produced for Day 1 is the lowest since raters spent some of the 6 hours in calibration. In addition, for most days of the study, fewer argument scores were produced than were issue scores. Thus it appears that, overall, argument prompts take longer to score than issue prompts.

Agreement rates overall appear to stay relatively steady (mid to high 70%) until Day 4, when there is a reduction in the percentage of raters with exact agreement. This was seen for both prompt types. On Day 5, there is a slight upturn in percentages for both prompt types, but these rates were not as high as those seen in Days 1–3.

This gradual change in exact agreement rates is also evident by examining contingency tables for each study day (see Appendix B). As the study day moves further away from calibration day (Day 1), the values on the diagonal tend to decrease and those on the off-diagonal increase. For example, for argument at a score level of 4, the percentage of raters giving the same score as the exemplar essay score decreases from 68% on Day 1 to 54% on Day 4 and 62% on Day 5. The same pattern is seen for issue, with percentages at a score level of 4 decreasing from 73% on Day 1 to 54% on Day 4 and 61% on Day 5.

Table 1 also provides the average Cohen's kappa across all raters for each of the study days by prompt type. Kappas range from −1 to 1, where $K = 1$ indicates perfect agreement and $K = 0$ indicates no agreement. At ETS, $K = .70$ is used as an evaluation threshold; values lower than .70 are considered unacceptable. For both prompt types, acceptable kappa values were seen on Days 1 and 2 and for Day 3 for issue. The value for Day 3 for argument was just under an acceptable threshold. However, for Days 4 and 5, for both prompts, the values are in the .60s, indicating that interrater agreement rates were below acceptable levels on those days, especially compared to the previous 3 days. In addition, as displayed in Table 2, the QWK values were extremely high. These results are most likely the result of the uniform distribution of the scores associated with the exemplar essays, so that the equal proportion of each score level is not representative of what is normally encountered in operational practice. The QWK also accounts for score adjacencies and is influenced by the very high exact + adjacent agreement rates. While QWK values are reported in Table 1, we believe that the exact agreement rates, kappa, and the contingency tables in Appendix B are more informative measures of rater performance here.

Comparisons in the exact agreement rates across the 5 study days were then examined. First, performance on one day was compared to performance on the next day using the mean exact agreement rate for all essays. This demonstrates changes in performance levels on a day-to-day basis and indicates if performance changed from the preceding day. In all cases, changes in the exact agreement rate were very small (see Table 2).

Because the goal of the study was to understand the role of calibration, differences in exact agreement rates were also examined using Day 1 (calibration day) as the baseline. This comparison allowed us to determine if performance was different on noncalibration days (Days 2, 3, 4, and 5) versus calibration day, even if performance on a particular day was not different than performance on the preceding day. Again, in all cases, changes in the exact agreement rate were very

Table 1 Mean Agreement Rates and Kappas by Study Day and Prompt Type Using All Essays

Prompt type	Study day	No. scores	Exact agreement	Kappa	QWK
Argument	1	1,747	.78 (.41)	.74	.96
	2	2,242	.76 (.43)	.71	.95
	3	2,134	.74 (.44)	.69	.94
	4	2,218	.71 (.45)	.65	.94
	5	2,004	.73 (.44)	.67	.94
Issue	1	1,823	.77 (.42)	.73	.95
	2	2,421	.75 (.43)	.70	.95
	3	2,408	.75 (.43)	.70	.95
	4	2,430	.71 (.45)	.65	.94
	5	2,328	.73 (.44)	.68	.94

Note. QWK = quadratic weighted kappa.

Table 2 Day-to-Day Changes in Exact Agreement Rates for All Essays

Prompt type	Comparison days	Change in exact agreement rate	Standard error
Argument	Day 1 to Day 2	.02	.02
	Day 2 to Day 3	.03	.02
	Day 3 to Day 4	.03	.01
	Day 4 to Day 5	−.03	.01
Issue	Day 1 to Day 2	.02	.01
	Day 2 to Day 3	.01	.01
	Day 3 to Day 4	.05	.01
	Day 4 to Day 5	−.03	.02

Table 3 Changes in Exact Agreement Rates Using Day 1 as Baseline for All Essays

Prompt type	Comparison days	Change in exact agreement rate	Standard error
Argument	Day 1 to Day 2	−.02	.02
	Day 1 to Day 3	−.06	.02
	Day 1 to Day 4	−.08	.02
	Day 1 to Day 5	−.06	.02
Issue	Day 1 to Day 2	−.02	.01
	Day 1 to Day 3	−.02	.01
	Day 1 to Day 4	−.07	.01
	Day 1 to Day 5	−.04	.01

small. As seen in Table 3, agreement rates on Day 1 were better than those on Days 3, 4, and 5 for argument and Days 4 and 5 for issue.

As described previously, to account for the potential variation in essay difficulty and differences in the number of essays read by raters, modified agreement rates were calculated using only those essays that had operational agreement rates. Table 4 summarizes this information.

As can be seen in this case, the agreement rates based on only those essays with operational data over the 5 study days are similar in magnitude (e.g., in the .70s) as those seen in Table 1 for both prompt types. The modified agreement rates are steady across the 5 days for argument but show a decline starting on Day 2 for issue. In addition, the increase in mean agreement rates seen on Day 5 does not occur when the modified agreement rates are used.

Table 4 also provides Cohen's kappa and QWK for each of the study days by prompt type for the subset of essays for which operational agreement data were available. For both prompt types, acceptable kappa values were seen on Days 1, 2, and 3 for argument and Days 1, 3, and 5 for issue, with the value for Day 2 for issue just under an acceptable threshold. However, for Days 4 and 5 for argument and Day 4 for issue, the values are in the .60s, indicating that interrater agreement rates were lower on those days.

Table 4 Kappas and Agreement Rates by Study Day and Prompt Type Using Only Essays With Operational Agreement Rates

Prompt type	Study day	No. of scores	Kappa	QWK	Modified agreement rate[a]	Operational agreement rate	Difference
Argument	Day 1	696	.74	.96	.79	.83	−.04
	Day 2	904	.71	.95	.76	.80	−.04
	Day 3	949	.70	.94	.75	.79	−.04
	Day 4	1,159	.67	.94	.73	.78	−.05
	Day 5	794	.65	.95	.71	.76	−.05
Issue	Day 1	665	.79	.96	.83	.83	.00
	Day 2	1,111	.69	.95	.75	.79	−.04
	Day 3	1,299	.73	.95	.78	.83	−.06
	Day 4	1,068	.66	.94	.72	.79	−.08
	Day 5	1,056	.65	.93	.71	.80	−.09

Note. QWK = quadratic weighted kappa.
[a] Modified agreement rates include only study essays that had operational agreement rates.

Table 5 Day-to-Day Changes in Modified Exact Agreement Rates

Prompt type	Comparison days	Change in exact agreement rate	Standard error
Argument	Day 1 to Day 2	.004	.02
	Day 2 to Day 3	−.003	.02
	Day 3 to Day 4	−.005	.02
	Day 4 to Day 5	.002	.02
Issue	Day 1 to Day 2	−.043	.02
	Day 2 to Day 3	−.011	.02
	Day 3 to Day 4	−.021	.02
	Day 4 to Day 5	−.007	.02

*Modified agreement rates include only study essays that had operational agreement rates.

Table 6 Changes in Modified Exact Agreement Rates Using Day 1 as Baseline

Prompt type	Comparison days	Change in exact agreement rate	Standard error
Argument	Day 1 to Day 2	.004	.02
	Day 1 to Day 3	.001	.02
	Day 1 to Day 4	−.005	.02
	Day 1 to Day 5	−.003	.02
Issue	Day 1 to Day 2	−.043	.02
	Day 1 to Day 3	−.053	.02
	Day 1 to Day 4	−.074	.02
	Day 1 to Day 5	−.081	.02

*Modified agreement rates include only study essays that had operational agreement rates.

As discussed previously, the QWK values were again extremely high. Thus, while these values are also reported in Table 4 for the reader's information, exact agreement rates and kappa act as more appropriate measures of rater performance in this context, and the QWK results are not discussed further.

As done previously, performance on one day was first compared to performance on the next day using the modified agreement rate. Only very small changes in agreement rates are apparent, as seen in Table 5. This indicates that when the operational agreement rate is taken into account, the raters in the study performed very similarly across all 5 study days for argument. For issue, there is a decline in exact agreement rates between Day 1 and Day 2, indicating that raters performed slightly less well on Day 2.

Differences in the modified agreement rates were also examined using Day 1 (calibration day) as the baseline. As seen in Table 6, differences between the agreement rates on noncalibration days (Days 2, 3, 4, and 5) were very small for argument compared to those seen on calibration day (Day 1). For issue, however, there were noticeable differences between Day 1 and Days 3, 4, and 5, with a reduction in agreement rates on those days. This indicates that when the operational agreement

Table 7 Mean Modified and Operational Exact Agreement Rates for Argument Prompt by Score Level for Each Study Day

Score level	Study day	No. of scores	Modified agreement rate[a]	Operational agreement rate	Difference
1	1	137	.95	.98	−.03
	2	138	.80	.94	−.14
	3	99	.80	.88	−.08
	4	138	.68	.83	−.15
	5	121	.72	.78	−.06
2	1	119	.91	.94	−.04
	2	118	.86	.85	.00
	3	145	.87	.84	.03
	4	184	.90	.89	.01
	5	127	.87	.87	.01
3	1	127	.84	.91	−.07
	2	133	.86	.89	−.04
	3	166	.81	.80	.01
	4	262	.79	.81	−.02
	5	96	.79	.84	−.05
4	1	108	.86	.86	.00
	2	183	.83	.81	.02
	3	214	.89	.83	.06
	4	197	.86	.81	.05
	5	109	.90	.82	.07
5	1	86	.56	.62	−.06
	2	151	.62	.66	−.03
	3	149	.54	.71	−.16
	4	242	.51	.62	−.11
	5	183	.57	.67	−.10
6	1	119	.51	.58	−.07
	2	181	.63	.69	−.06
	3	176	.57	.72	−.15
	4	136	.63	.76	−.13
	5	158	.56	.67	−.11

[a] Modified agreement rates include only study essays that had operational agreement rates.

rate is taken into account, the raters in the study performed very similarly across all 5 study days on the argument prompt but not the issue prompt, suggesting that accuracy is moderated by prompt type.

The mean modified and operational agreement rates were also examined at the score level for both the study and operational agreement rates for the subset of responses for which operational agreement data were available. There were two reasons for this. First, it was observed that, overall, exemplar essays representing the very bottom (1s) and top (6s) of the scale had lower agreement rates when used during operational scoring. Second, the overrepresentation of 1s and 6s may have resulted in agreement rates appearing to be lower than they might have been if a more realistic proportion of these essays had been used in the study. Table 7 presents this information for the argument prompt, and Table 8 presents this information for the issue prompt.

For most days, the difference between the modified agreement rates and operational agreement rates were small. However, somewhat larger differences are seen with the 1s and 6s on Days 3, 4, and 5. This was apparent for both prompt types, but especially for issue.

Rater Survey

A total of 32 of the 46 raters voluntarily completed the short survey sent to them following the end of the scoring study. Table 9 presents the number of raters responding to each survey question option.

Most raters in the study were very experienced raters, with the highest percentage of raters having scored operationally for more than 5 years. This reflects the characteristics of the GRE rater pool, where few new raters are added yearly. The majority of the raters also were scheduled for 8-hour operational scoring shifts, again reflective of operational scoring for GRE.

Table 8 Adjusted Mean Modified and Operational Exact Agreement Rates for Issue Prompt by Score Level for Each Study Day

Score level	Study day	No. of scores	Modified agreement rate[a]	Operational agreement rate	Difference
1	1	54	.89	.93	−.04
	2	135	.86	.95	−.09
	3	187	.80	.85	−.05
	4	136	.86	.90	−.04
	5	153	.78	.90	−.12
2	1	134	.73	.73	.00
	2	177	.76	.78	−.03
	3	216	.85	.86	−.01
	4	184	.82	.82	.00
	5	174	.68	.80	−.12
3	1	129	.96	.91	.05
	2	176	.80	.81	−.01
	3	266	.84	.86	−.02
	4	186	.84	.85	−.01
	5	173	.91	.87	.05
4	1	125	.96	.93	.03
	2	244	.84	.80	.04
	3	263	.87	.85	.02
	4	152	.81	.70	.11
	5	214	.80	.80	.00
5	1	130	.68	.72	−.04
	2	201	.60	.75	−.15
	3	196	.54	.72	−.18
	4	229	.43	.66	−.24
	5	214	.50	.67	−.17
6	1	93	.77	.82	−.05
	2	178	.65	.70	−.05
	3	171	.69	.85	−.16
	4	181	.67	.88	−.21
	5	128	.61	.78	−.17

[a] Modified agreement rates include only study essays that had operational agreement rates.

Most raters indicated that they felt their scoring behavior remained relatively stable across all 5 study days. The majority of raters had the same level of confidence in assigning scores across all study days, felt they scored at the same speed each day, and referred to benchmarks with the same frequency. For those raters who did not feel their behavior was consistent across all 5 days, however, some interesting results emerged. For example, more of these raters felt more confident when assigning scores in later days than in earlier days, with the percentage of raters being similar across both types of prompts. However, for scoring pace, a greater percentage of raters who scored the argument prompt believed they scored faster in the earlier days, while a greater percentage of raters who scored the issue prompt believed they scored faster in the later days. Differences were also seen for the use of benchmarks: Those raters who scored the argument prompt referred to benchmarks more frequently in later days, while those who scored the issue prompt referred to benchmarks more frequently in the earlier days.

Finally, when asked what the ideal length for a GRE scoring shift would be, the majority of raters (57%) indicated 6 hours. Approximately one quarter of the raters (23%) indicated that 8 hours would be the ideal scoring shift.

Conclusions and Discussion

Although past studies have indicated the importance in calibration as part of scoring accuracy, the question remains as to how frequently it is needed to ensure that raters continue to know and accurately apply the scoring criteria. The current study attempted to address this question by allowing raters to continue scoring under controlled conditions for multiple days following a single calibration.

The design of the study was intended to mimic the operational scoring process used for GRE. Raters used the same platform (ONE) they use in operational scoring to read and assign scores as part of the study; as occurs operationally,

Table 9 Number of Raters Responding to Each Survey Question Option

Survey question	Response option	All raters	Argument raters	Issue raters
Experience working as GRE rater	<1 year	9	6	2
	1–5 years	3	2	1
	>5 years	20	9	11
Usual shift length	4 hours	8	4	3
	8 hours	22	12	10
	Other	1	0	1
Confidence level assigning scores	More confident in earlier days than in later days	3	2	1
	More confident in later days than in earlier days	9	5	3
	Same level of confidence during all days	20	10	10
Scoring pace	Scored faster in earlier days than in later days	5	5	0
	Scored faster in later days than in earlier days	6	3	3
	Scored at same pace during all days	21	9	11
Use of benchmarks	Used more frequently in earlier days than in later days	8	3	4
	Used more frequently in later days than in earlier days	5	4	1
	Same frequency during all days	18	10	8
	Never reviewed benchmarks	1	0	1
Ideal scoring shift	4 hours	5	3	2
	6 hours	17	8	9
	8 hours	7	4	3
	Other	1	1	0

Note. Because the survey was anonymous, raters who did not indicate the type of prompt they scored only appear in the all raters column. Thus all raters may not equal the total of argument raters plus issue raters.

raters were assigned a single prompt type (argument or issue) to score over all scoring days; and raters had access to the same scoring tools (benchmarks) they have during operational scoring. It was hoped that these similarities would help mask the fact that the essays raters were scoring were not operational and encourage them to score with the same care and thoughtfulness used when scoring operationally.

However, there were differences in the study that may have influenced rater behavior. First, raters were allowed to score only for 6 hours each day and only 3 hours per prompt. In operational scoring, raters generally score for 8 hours. For the raters in this study, this was definitely the case, with 22 of the 31 raters responding to the survey question indicating their usual scoring shift consisted of 8 hours.

Second, the score points represented by the essays used in the study did not accurately reflect what raters would expect to see operationally. A comparison of operational agreement rates at the score level indicated that while the differences between the study agreement rates and operational agreement rates were small overall, larger differences were seen with the 1s and 6s on Days 3, 4, and 5. The disproportionate number of essays representing the top and bottom ends of the score scale may have artificially decreased the exact agreement rates. This decrease could also explain the difference in findings between this study and the Wendler and Cline (2017) study, where the distribution of score points was more typical than observed here.

How does calibration impact performance on subsequent days? For both argument and issue, exact agreement rates were highest on the day that raters calibrated (Day 1). Interrater agreement, as measured by Cohen's kappa, was also highest for Day 1. In addition, the lowest kappas were seen for Days 4 and 5 for both prompt types.

When comparing agreement rates for Days 2, 3, 4, and 5 with the agreement rate seen with Day 1 (calibration day), a different picture emerged. When all essays were included in the analyses, agreement rates for Days 3, 4, and 5 for argument and Days 4 and 5 for issue were different than those seen on Day 1. When comparisons were done using only those prompts with known operational agreement rates, however, there were no differences in agreement rates between Day 1 and the subsequent days for argument. For issue, Days 3, 4, and 5 were different from Day 1. However, the difference seen on Day 3 compared to Day 1 for issue may be a spurious result in that, regardless of whether the analysis was based on the agreement rate using all essays or the rate based on the subset of essays with known operational agreement rates, the agreement rates were very similar and the kappas were at or above the acceptable threshold.

The results of these analyses point to the role that calibration plays in ensuring that raters appropriately and consistently apply the scoring rubrics. While most raters believed that their scoring performance in terms of accuracy and speed did not change, results indicate that this was not necessarily so. There were only small differences in exact agreement rates between Days 1, 2, and 3 for both prompts but slightly larger differences in later days. This suggests that the influence of calibration remains with raters through at least the next couple of scoring days. The comparison with modified mean differences also suggests that for argument, there is minimal to no decline over the 5 days, whereas for issue, there appears to be more of a steady decline. This indicates that change in accuracy is moderated by prompt type and any change in calibration frequency needs to take into account the prompt–accuracy interaction.

It should be noted that the agreement rates were based on the most conservative index: exact agreement. In practice, rater performance is monitored by percentage of exact + adjacent scores as well as the number of discrepant scores. Because the exact + adjacent agreement rate was so high, it was decided to use exact scores only to determine agreement rates. In addition, the equal proportion of essays across all score levels undoubtedly had an impact on agreement rates. This impact became clear when comparisons were made by score level.

Even though the distribution of score points represented by the essays in this study does not necessarily represent what raters generally see during operational scoring, the preceding results indicate that daily calibration may not be necessary to ensure scoring quality. However, it appears that the impact of calibration may be moderated by the type of prompt being scored. It is not known if the results from this study would generalize to other prompt or task types.

From the results of this study, it appears that requiring raters to calibrate every third day (e.g., Day 1, Day 4, Day 7) of scoring is unlikely to impact rater scoring accuracy for either the issue or argument prompt. The results further suggest that even less frequent calibration for raters scoring the argument prompt would not lead to diminished scoring quality. However, prompt type did appear to moderate accuracy, and given the decline in performance that occurred on Day 4 and 5 for issue, not requiring calibration every third day could result in a decrease in scoring accuracy for that prompt type. Therefore, the calibration schedule for all raters should be based on a conservative approach. It must take into account the differences seen between the two types of prompts and be set to reflect the needs of the issue task. In addition, because raters do not know which prompt type they will be assigned, having inconsistent calibration rules will likely prove to be confusing to them as part of operational scoring.

Calibration functions as a “gatekeeper” to operational scoring. As a result, raters spend more time scoring individual essays during calibration than they do under an operational setting (Finn, Wendler, Ricker-Pedley, & Arslan, 2018). This is not unexpected, in that unless raters pass calibration, they cannot move on to operational scoring and thus lose wages. As one rater commented in the rater survey, “It was . . . great to only have to calibrate on the first day. . . . I truly felt that not having to calibrate made me relax more during the shift.” Revising the policy for less frequent calibration will not only result in more productive scoring time but would also likely be viewed positively by raters.

While the results of this study indicate the role of calibration in ensuring accurate rater performance, additional studies are warranted. The results of this study indicate that rater performance may be impacted by prompt type, and therefore studies examining calibration with other prompt or tasks types are needed. Although it was not possible to examine individual rater characteristics in this study, future studies should consider evaluating variables like amount of prior rating experience, type of training, and other background variables in addition to examining different prompt or task types. Finally, in addition to Cohen’s kappa, future studies could also consider examining rater consistency using other indices of interrater agreement or interrater reliability, such as joint probability of agreement, interclass correlations, or other correlation coefficients.

Acknowledgments

The authors acknowledge and thank the following individuals for their work on this project: J. Wyman Brantley, Heather Fell, Diane Pierce, Sara Stern, Gail Strelko, Valerie Thompson, and Scott Zimmermann.

Notes

[1] At the time of the study, the GRE rater pool consisted of 250–300 raters. Thus the number of raters used in the study represented about 15%–18% of the available pool.

[2] Scoring sessions for most GRE raters consist of either 4-hour or 8-hour shifts.

[3] The remaining essays had fewer than 20 raters who had scored them.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *31*, 99–115. https://doi.org/10.1177/0265532215582283

Baldwin, D., Fowles, M., & Livingston, S. (2008). *Guidelines for constructed-response and other performance assessments*. Princeton, NJ: Educational Testing Service.

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

Bridgeman, B., Trapani, C., & Bivens-Tatum, J. (2011). Comparability of essay question variants. *Assessing Writing, 16*, 237–255. https://doi.org/10.1016/j.asw.2011.06.002

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

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*, 163–178. https://doi.org/10.1111/j.1745-3984.2000.tb01081.x

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

Everson, P. & Hines, S. (2010). How ETS scores the TOEIC® Speaking and Writing Test responses. In D. E. Powers (Ed.), *The Research Foundation for the TOEIC® Tests. A compendium of studies* (pp. 8.1–8.9). Princeton, NJ: ETS.

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?* (Research Report No. RR-18-31). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12217

Glazer, N. (2017, April). *The rater calibration process*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–432). Westport, CT: American Council on Education/Praeger.

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

McClellan, C. (2010, February). Constructed-response scoring: Doing it right. *R&D Connections*, no. 13.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*, 371–389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

Parke, C. S., Lane, S., & Stone, C. A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation*, *12*, 239–269. https://doi.org/10.1080/13803610600696957

Ricker-Pedley, K. (2011). *An examination of the link between rater calibration performance and subsequent scoring accuracy in Graduate Record Examinations® (GRE®) Writing* (Research Report No. RR-11-03). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02239.x

Ricker-Pedley, K. L., & Li, H. (2010). Rater calibration and subsequent scoring performance [Internal manuscript]. Princeton, NJ: Educational Testing Service.

Wendler, C., & Cline, F. (2017, April). *Rater scoring accuracy across a multiple-day scoring window*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*, 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

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

Zhang, M. (2013, March). Contrasting automated and human scoring of essays. *R&D Connections*, no. 21.

Appendix A

Rater Survey

1. **How long have you worked as a GRE rater?**

□ Less than one year
□ One to five years
□ More than five years

2. **What shift length do you generally work as a GRE rater?**

□ Four hours
□ Eight hours
□ Other

3. **Which best describes your confidence with assigning scores during the essay scoring project?**

□ I felt more confident assigning scores in the **earlier** days than in the **later** days.
□ I felt more confident assigning scores in the **later** days than in the **earlier** days.
□ I felt the same level of confidence assigning scores during all days of scoring.

4. **Which best describes your scoring pace during the essay scoring project?**

□ I felt I scored faster in the **earlier** days than in the **later** days.
□ I felt I scored faster in the **later** days than in the **earlier** days.
□ I felt I scored at the same pace during all days of scoring.

5. **Which best describes your use of benchmarks during the essay scoring project?**

□ I reviewed the benchmarks more frequently in the **earlier** days than in the **later** days.
□ I reviewed the benchmarks more frequently in the **later** days than in the **earlier** days.
□ I reviewed the benchmarks with the same frequency during all days of scoring.
□ I never reviewed the benchmarks.

6. **Which do you believe is the most ideal length for a GRE scoring shift?**

□ Four hours
□ Six hours
□ Eight hours
□ Other

Appendix B

Contingency Tables

Table B1 Contingency Table for Argument Prompt for Study Day 1

	Exemplar essay score						
	1	2	3	4	5	6	Total
Rater score							
1	**260**	16	0	0	0	0	276
2	11	**250**	33	0	0	0	294
3	0	22	**248**	29	3	0	302
4	0	1	23	**258**	94	6	382
5	0	0	0	13	**178**	106	297
6	0	0	0	1	19	**176**	196
Total	271	289	304	301	294	288	1, 747

Note. Numbers in bold are the number of raters having exact agreement with the exemplar essay score.

Table B2 Contingency Table for Argument Prompt Study for Day 2

	Exemplar essay score						
	1	2	3	4	5	6	Total
Rater score							
1	**276**	10	0	0	0	0	286
2	73	**286**	14	0	0	0	373
3	1	48	**347**	44	2	0	442
4	0	6	34	**324**	139	9	512
5	0	1	0	14	**225**	128	368
6	0	0	0	0	17	**243**	260
Total	350	351	395	382	383	380	2, 242

Note. Numbers in bold are the number of raters having exact agreement with the exemplar essay score.

Table B3 Contingency Table for Argument Prompt for Study Day 3

	Exemplar essay score						
	1	2	3	4	5	6	Total
Rater score							
1	**252**	21	0	0	0	0	273
2	84	**298**	26	0	0	0	408
3	1	25	**310**	37	1	0	374
4	0	4	39	**331**	136	16	526
5	0	2	1	7	**202**	140	352
6	0	1	0	0	14	**186**	201
Total	337	351	376	375	353	342	2, 134

Note. Numbers in bold are the number of raters having exact agreement with the exemplar essay score.

Table B4 Contingency Table for Argument Prompt for Study Day 4

	Exemplar essay score						
	1	2	3	4	5	6	Total
Rater score							
1	**251**	10	0	0	0	0	261
2	97	**325**	35	0	0	0	457
3	0	31	**321**	60	2	0	414
4	0	2	57	**287**	172	13	531
5	0	0	1	10	**234**	136	381
6	0	0	0	0	22	**152**	174
Total	348	368	414	357	430	301	2, 218

Note. Numbers in bold are the number of raters having exact agreement with the exemplar essay score.

Table B5 Contingency Table for Argument Prompt for Study Day 5

	Exemplar essay score						
	1	2	3	4	5	6	Total
Rater score							
1	**229**	22	0	0	0	0	251
2	93	**274**	15	0	0	0	382
3	0	21	**271**	17	1	0	310
4	0	1	39	**303**	142	7	492
5	0	0	3	26	**200**	138	367
6	0	0	0	0	18	**184**	202
Total	322	318	328	346	361	329	2, 004

Note. Numbers in bold are the number of raters having exact agreement with the exemplar essay score.

Table B6 Contingency Table for Issue Prompt for Study Day 1

	Exemplar essay score						
	1	2	3	4	5	6	Total
Rater score							
1	**203**	11	0	0	0	0	214
2	54	**211**	22	0	0	0	287
3	2	53	**282**	17	0	0	354
4	0	0	19	**329**	98	7	453
5	0	0	4	10	**198**	95	307
6	0	0	0	0	19	**189**	208
Total	259	275	327	356	315	291	1, 823

Note. Numbers in bold are the number of raters having exact agreement with the exemplar essay score.

Table B7 Contingency Table for Issue Prompt for Study Day 2

	Exemplar essay score						
	1	2	3	4	5	6	Total
Rater score							
1	**269**	22	0	0	0	0	291
2	52	**321**	43	0	0	0	416
3	1	62	**379**	47	1	0	490
4	0	0	25	**375**	147	7	554
5	0	0	0	31	**236**	122	389
6	0	0	0	0	37	**244**	281
Total	322	405	447	453	421	373	2, 421

Note. Numbers in bold are the number of raters having exact agreement with the exemplar essay score.

Table B8 Contingency Table for Issue Prompt for Study Day 3

	Exemplar essay score						
	1	2	3	4	5	6	Total
Rater score							
1	**306**	13	0	0	0	0	319
2	74	**331**	29	0	0	0	434
3	0	53	**360**	30	2	0	445
4	0	1	50	**395**	166	6	618
5	0	0	0	22	**188**	122	332
6	0	0	0	0	27	**233**	260
Total	380	398	439	447	383	361	2, 408

Note. Numbers in bold are the number of raters having exact agreement with the exemplar essay score.

Table B9 Contingency Table for Issue Prompt for Study Day 4

	Exemplar essay score						
	1	2	3	4	5	6	Total
Rater score							
1	**316**	6	0	0	0	0	322
2	67	**330**	33	0	0	0	430
3	3	77	**317**	77	4	1	479
4	0	0	37	**309**	221	7	574
5	0	0	1	12	**193**	140	346
6	0	0	0	0	20	**259**	279
Total	386	413	388	398	438	407	2, 430

Note. Numbers in bold are the number of raters having exact agreement with the exemplar essay score.

Table B10 Contingency Table for Issue Prompt for Study Day 5

	Exemplar essay score						
	1	2	3	4	5	6	Total
Rater score							
1	**269**	5	0	0	0	0	274
2	88	**295**	13	0	0	0	396
3	6	67	**382**	52	6	1	514
4	0	1	34	**346**	174	9	564
5	0	0	1	29	**198**	112	340
6	0	0	0	0	29	**211**	240
Total	363	368	430	427	407	333	2, 328

Note. Numbers in bold are the number of raters having exact agreement with the exemplar essay score.

Suggested citation:

Wendler, C., Glazer, N., & Cline, F. (2019). *Examining the calibration process for raters of the GRE general test* (GRE Board Research Report No. GRE-19-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12245

Action Editor: Brent Bridgeman

Reviewers: This report was reviewed by the GRE Technical Advisory Committee and the Research Committee and Diversity, Equity and Inclusion Committee of the GRE Board.