Research Report

Does the Time Between Scoring Sessions Impact Scoring Accuracy? An Evaluation of Constructed-Response Essay Responses on the *GRE*® General Test

ETS RR–18-31

Bridgid Finn
Cathy Wendler
Kathryn L. Ricker-Pedley
Burcu Arslan

December 2018

ETS Research Report Series

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

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

RESEARCH REPORT

Does the Time Between Scoring Sessions Impact Scoring Accuracy? An Evaluation of Constructed-Response Essay Responses on the *GRE*® General Test

Bridgid Finn, Cathy Wendler, Kathryn L. Ricker-Pedley, & Burcu Arslan

Educational Testing Service, Princeton, NJ

This report investigates whether the time between scoring sessions has an influence on operational and nonoperational scoring accuracy. The study evaluates raters' scoring accuracy on constructed-response essay responses for the *GRE*® General Test. Binomial linear mixed-effect models are presented that evaluate how the effect of various predictors, such as time spent scoring each response, days in a scoring gap, and number of consecutive days of scoring, relate to scoring accuracy. Results suggest that for operational scoring, the number of days in a scoring gap has a negative influence on performance. The findings, as well as other results from the models, are discussed in the context of cognitive influences on knowledge and skill retention.

Keywords *GRE*® General Test; CR scoring; forgetting over time

doi:10.1002/ets2.12217

Complex performance tasks, such as written essay and verbal responses (i.e., constructed-response [CR] items), are increasingly being used in numerous standardized assessment programs. A number of well-known tests, such as the Graduate Management Admission Test (GMAT), the *SAT*® test, the *GRE*® General Test, and many K-12 state assessments, include at least one essay-writing item (Zhang, 2013). Other tests, such as the *TOEFL iBT*® test and the Pearson Test of English, also include CR items in the form of spoken responses. CR items and tasks are used to measure knowledge and skills that are not as adequately evaluated using multiple-choice questions (Braun, 1988). For example, essay responses typically integrate a student's skill in reading comprehension, text analysis, and writing skills (for a review, see Deane, 2011).

Human raters typically score CR responses. These raters vary in the total number of days per year that they are scheduled to score as well as in their scoring schedules. For example, raters differ in the number of consecutive days of scoring that they may do and in the gap between their scoring sessions. The purpose of this study was to examine how the time between scoring sessions impacted nonoperational and operational scoring accuracy of written essay responses. We approached this topic by considering how human cognition influences raters' CR scoring performance. Specifically, we anticipated that gaps between scoring sessions would lead to forgetting and skill decline, which would negatively impact performance. The findings will be discussed from the perspective of cognitive psychology and will highlight the impact of time between sessions on the retention of knowledge and skills.

The Impact of Human Raters on CR Scoring

Human Error

Human raters, who understand the content and quality of writing, are primary in evaluating essay responses. To assign scores, raters are trained to utilize rubrics that specify characteristics of essays representing the score levels. Unfortunately, however, human raters can introduce error into the scoring process, which can weaken scoring reliability. Applying scoring rubrics requires some level of human interpretation and subjective judgment, which can differ from rater to rater, and even moment to moment for a rater, even with careful training. According to Guilford (1936), "raters are human and they are therefore subject to all the errors to which humankind must plead guilty" (p. 272). These errors may manifest as

Corresponding author: B. Finn, E-mail: bfinn@ets.org

inter- and intra-rater agreement inconsistencies and can result in severity biases, central tendency effects, halo effects, and restricted range effects, to name a few (Engelhard, 1994; Ricker-Pedley, 2011; Saal, Downey, & Lahey, 1980; Zhang, 2013). For example, a rater's severity, or strictness in scoring, can change over the course of their scoring session (e.g., Braun, 1988; Hoskens & Wilson, 2001). Severity differences between raters also occur: One rater may tend to assign relatively high or lenient scores, whereas another might score with greater strictness (Coffman, 1971). For test takers taking tests that aid in making high-stakes decisions, there are important consequences of rater bias. Two students whose essays should receive the same score could potentially receive different scores depending on whether they had a more lenient or strict rater.

Calibration Practices

A number of monitoring practices, including calibration, back scoring, and double scoring, among others, are used to mitigate against human error. Calibration, a focus of the current study, is a widely used method used to control rater error and ensure the accuracy and reliability of human ratings. It is a practice undertaken before raters begin operational scoring (Congdon & McQueen, 2000; Lumley & McNamara, 1995). Rater calibration helps to ensure that the ratings produced remain consistent across and between scoring sessions and is a scoring practice commonly used in large-scale testing programs.

Calibration typically occurs at the start of a day of scoring, with a scoring shift, or when a new CR prompt is introduced. During calibration, a rater scores a short set of prescored "exemplar responses," which are responses that have previously been reviewed by experts and have been deemed as clear exemplars of the distinct score levels. The scores for the calibration samples have typically been determined by multiple independent ratings, followed by group consensus by scoring leaders to determine the "true" score. These samples are thus considered a gold standard by which to measure rater performance (Ricker-Pedley, 2011).

Prior to calibration, a rater can review training materials specific to the prompt that he or she will be scoring. These training materials commonly include the scoring rubric and benchmark essays, which provide detailed descriptions and examples of the performance that qualifies a response at each score level (Parke, Lane, & Stone, 2006). While calibration can be initiated in some instances without any review of the training materials, raters typically do review these training materials prior to the starting to score the samples if there has been a long gap since their last scoring session (Finn & Roth, 2017). A main benefit of training before calibration is to orient the rater to the rating scale and to improve intrarater consistency (Congdon & McQueen, 2000; Wigglesworth, 1994).

Calibration may be used as a criterion test in which a minimum performance standard must be met before a rater is allowed to move on to score operational samples. If, for example, a rater is not scoring enough samples in the calibration set as exact (exact match between rating and true score) and instead gives too many discrepant (off by more than 1 scale point) or adjacent (off by 1) scores, he or she will not pass that calibration attempt. The rater can, but does not necessarily, review training materials before his or her next calibration attempt. Raters are usually discharged from scoring for the day if they have failed two attempts at calibration (Ricker-Pedley, 2011). In sum, the calibration process serves as a type of quality control for the CR scoring process (McClellan, 2010; Myford & Wolfe, 2009).

Successfully completing calibration has been found to be predictive of accurate scoring on the same day that calibration occurs (Ricker-Pedley, 2011). For example, Ricker-Pedley (2011) conducted an experimental study to examine the link between calibration accuracy and subsequent operational scoring accuracy. Raters scored 75 calibration responses followed by 100 operational samples using prompts retired from the GRE General Test pool. The results demonstrated that even a 10-item GRE calibration test was predictive of subsequent operational scoring accuracy.

Though a link between calibration and operational scoring accuracy has been established, much less is known about how gaps between scoring sessions impact scoring accuracy. Consider a rater who scores less frequently, perhaps because of gaps in his or her scoring schedule. Will the rater's scoring performance be poorer than that of another rater who scores more regularly? A goal of the current study was to evaluate how the time between scoring sessions relates to nonoperational and operational scoring accuracy.

Scoring Frequency and Rater Cognition

To score an essay response, raters must draw on previously trained skills and knowledge. During scoring, raters evaluate the information provided in the essay, connect it with their own prior knowledge, appraise the factual correctness of claims made, determine the how the essay aligns with the rubric, and, finally, make a judgment about what score the essay merits (Zhang, 2013). Before scoring, a rater typically reviews training materials specific to the prompt that he or she will be scoring. These training materials commonly include the scoring rubric and benchmark essays, which provide detailed descriptions and examples of the performance that qualifies a response at each score level (Parke et al., 2006). While the specific prompt and benchmark essays included in the training set may change across scoring sessions, the underlying criterion separating score levels for a particular assessment does not. For example, the GRE includes two prompt types, issue and argument, both of which are scored using a 6-point holistic scale. An essay that merits a score of 3 (limited) must be distinguished from an essay that merits a 4 (adequate). Distinguishing essays across these score categories is a proficiency that raters likely develop with training and through scoring experience (but see Bond, 1995; Joe, Harmes, & Hickerson, 2011). The question we address here is whether gaps in calibration and scoring impact a rater's capacity to accurately make these distinctions and score correctly.

Other domains and disciplines have evaluated related questions, namely, how the time spent since training a particular skill or learning a set of knowledge influences long-term retention of skill or information (e.g., Arthur, Bennett, Stanush, & McNelly, 1998; Schmidt & Bjork, 1992). Ebbinghaus's (1964) pioneering experimental studies on forgetting (for a recent replication, see Murre & Dros, 2015) experimentally established the impairing effects of time on knowledge retention: The longer the retention interval (i.e., the period of time since training), the worse performance will be. In a meta-analysis of 189 studies, Arthur et al. (1998) found a substantial decay in skills and knowledge over time across a variety of domains. Arthur et al. defined skill decay as the "loss or decay of trained or acquired skills (or knowledge) after periods of nonuse" (p. 58). The review found that physical tasks, such as those that require muscular strength and coordination, showed less skill loss over time than cognitive tasks, such as problem solving and decision making.

Essay scoring, a cognitive task, was not evaluated in the review. However, given these findings, scoring performance would similarly be expected to decline with disuse. That is, a decline in scoring accuracy over time hypothesis would predict that as the gap between scoring sessions increases for a rater, scoring may become less accurate.

There are two important caveats to consider with respect to this prediction, however. First, gaps in a rater's scoring schedule are typically on the order of days and weeks rather than months and years, as were many of the retention intervals in the studies reported by Arthur et al. (1998). It is possible that a decline in essay scoring performance may not be in evidence over short scoring gaps (Myford & Wolfe, 2009). A second consideration is that raters generally are required to recalibrate at the start of each new scoring session. If they pass, they then move on to operational scoring. This calibration session may function similarly to a retraining session, during which time the rater will have had an opportunity to re-review the scoring rubric, benchmark essays, and practice scoring training essays. In addition, raters can always review the training materials, benchmarks, and scoring rubrics, which may lessen the impact of scoring gaps.

According to a calibration-as-retraining hypothesis, performance decrements may not be in evidence after a gap in scoring if calibration serves to effectively retrain raters' scoring skills. Indeed, decades of research in cognitive psychology have shown that spacing out (rather than blocked or consecutive) training episodes can be beneficial to learning and knowledge retention (e.g., Melton, 1967; Schmidt & Bjork, 1992). Thus spaced retraining may mitigate any loss in scoring skill that may have occurred over a retention gap.

Study Purpose

The purpose of this study was to examine how the time between scoring sessions impacted nonoperational and operational scoring accuracy. The time spent in calibration is considered nonoperational, that is, time in which the rater is not scoring operational samples. Though scoring accuracy standards are of course paramount, an important applied question is whether it is possible to reduce calibration frequency while maintaining the level of quality control that is in place with more frequent (i.e., daily) calibration. Large-scale scoring is labor intensive, time consuming, and expensive (Zhang, 2013). Thus reductions in nonoperational time could potentially reduce delays in score reporting to candidates and lower operational costs (Ricker-Pedley, 2011). We evaluated this question by examining whether there was a decline in scoring accuracy as the interval between scoring sessions increased.

Methods

For this study, we drew on data from a large CR data set extracted from the Online Network for Evaluation (ONE) system (a proprietary ETS system that enables raters to score responses to many types of CR tasks via secure Internet access). The data set included all scoring data from 2015 for ETS branded tests, including the GRE General Test, TOEFL iBT, the *TOEIC*® tests, and the *PRAXIS*® assessments. The current study focused on ratings from the GRE General Test and used data from 350 deidentified raters. The set included approximately 199 distinct prompts represented, with raters scoring 62 distinct prompts on average. The scores ranged from 1 to 4, with 1 representing the lowest score value and 4 representing the highest score value.

The 350 raters in the sample had previously been vetted through an initial screening process to ensure that they had adequate scoring credentials (i.e., adequate background expertise). In addition, before raters were hired for operational scoring, they were required to pass certification, a test given to raters following training that functions similarly to a calibration. All training, calibrating, and scoring was conducted using the ONE system so that raters scored at their own secure computer terminals at their home or workplace. Standard scoring instructions tell raters to score at their own pace.

Raters are required to pass calibration prior to advancing to operational scoring. Calibration sets comprise sample essays. To pass calibration and move on to operational scoring, GRE raters must have agreement rates of at least 60% exact agreement (defined as being exactly the same score as the predetermined score for a calibration paper) and 0% discrepant scores (defined as being 2 or more score points from the predetermined score). The established exact agreement rate of 60% represents the minimal level that is believed to be needed to ensure that a rater is appropriately applying the scoring rubrics; a rater who receives 90% exact agreement is not considered to be a better rater than one who receives 70% exact agreement. If a rater fails his or her first calibration attempt, he or she is given the opportunity to take a second calibration test. If the rater fails the second calibration test, he or she is released from scoring for the day. Immediately following successful calibration, raters begin to score operational samples.

The GRE was selected for evaluation of our research questions for a number of reasons. First, the size of the scoring pool made it more likely that we would find a sufficient number of raters with a variable range of days spent scoring and days spent without scoring (i.e., days in a gap). Second, the GRE writing prompts are scored on a 6-point scale, yielding adequate variability in score assignment.

There were no restrictions on the number of days that raters could score for inclusion in the analysis set. As can be seen in Table 1, there was a wide range in the number of days that raters scored in 2015. Note that the total number of days spent scoring and spent in a scoring gap does not sum to 365 because raters may have started scoring at any point during 2015. In addition, the scoring volume is seasonal. Volume is high in spring and lower in summer, which means that fewer raters are needed during particular months.

Analyses

Measures of Scoring Accuracy

We chose to evaluate the questions of how time between scoring sessions influenced accuracy by evaluating both nonoperational and operational scoring performance. Nonoperational performance was measured as calibration performance, whereas operational performance was measured as performance on validity samples embedded in the operational scoring session. Validity samples are similar to calibration samples in that they have been prescored by scoring experts. As with calibration samples, they are intended to represent clear examples of a particular score. Validity samples are "seeded" into test-taker responses being scored and, as such, may be used as a benchmark by which to assess rater performance (McClellan, 2010). Raters do not know which responses are validity samples and which are test takers' responses.

Table 1 Descriptive Statistics Days Spent in Scoring and Days Spent in a Scoring Gap

Variable and statistic	Mean (SE)	Min.	Max.
Total number of days in scoring	67.21 (2.47)	2	212
Total number of days spent in a scoring gap	244.45 (4.41)	5	342

Note. $n = 350$.

Table 2 Sample Scoring Profile of Two GRE Raters

Rater	Scoring day in year[a]	Days in scoring gap	Days of consecutive scoring	% Exact (calibration)	Average response time (calibration) (s)	% Exact (validity)	Average response time (validity)
1	10	–	1	.80	151	1.00	356
1	17	2	1	.90	222	.57	353
1	18	0	2	.90	230	.70	388
2	5	–	1	.70	522	.78	351
2	9	3	1	.70	292	.63	433
2	14	4	1	1.00	391	.71	528

[a] From 1 to 365.

Scoring accuracy was measured using exact agreement on both calibration samples and validity samples. Exact agreement referred to whether a rater had assigned a response (calibration or validity response) a score that was the same as the "true score." Other measures of rater accuracy, such as adjacent agreement, discrepant agreement, and rater bias, were not used in the current analysis. Adjacent agreement can be calculated by comparing whether the rater assigned a response a score that was either 1 point higher or lower than the assigned true score. Discrepant agreement can be calculated by comparing whether the rater assigned a response a score that was more than 1 point higher or lower than the assigned true score. Rater bias can be evaluated by computing the directional difference between the assigned calibration or validity score and the rater score.[1]

Design

Our goal was to evaluate how the time between scoring sessions and the days of consecutive scoring influenced scoring accuracy. To accomplish this, a scoring profile was developed for each rater, and we modeled the contribution of the gap between scoring sessions to predict whether raters assigned responses an exact score. Table 2 presents a subset of a sample scoring profile for two GRE raters. The profile provides an example of the type of variability in scoring schedules (days of consecutive scoring, days in gap) as well as some of the nonoperational (percentage exact for calibration) and operational (percentage exact validity) scoring metrics available for analyses.

Results

Table 3 reports participant means as they relate to the gaps between scoring sessions: total count of days in 2015 spent scoring, total count of days in scoring gaps (days in which no scoring occurred), average scoring gap, and days of consecutive scoring. Table 3 also includes the average seconds of time spent scoring each response and the scoring metrics for both calibration and validity samples each day. These descriptive statistics are followed by analyses relating time spent scoring days between scoring sessions to nonoperational and operational scoring metrics.

The decline in scoring accuracy over time hypothesis would predict that a longer time spent without scoring will lead to a drop in scoring accuracy. We evaluated this hypothesis with both calibration scoring accuracy and validity scoring accuracy. We modeled the contribution of the gap between scoring sessions as well as other variables to predict whether raters assigned responses an exact score. Though raters did have an opportunity to review rubrics and benchmark responses before they engaged in a calibration session, the calibration test itself is the first measure of performance since the previous scoring session. Validity scoring follows at least one round of calibration and occurs during operational scoring itself, which may allow suitable practice for recovering any scoring accuracy declines. Thus calibration accuracy may allow a clearer view of how the gap between scoring sessions and days of consecutive scoring influences scoring accuracy. Use of the validity samples offered an analysis of scoring performance as it relates to scoring accuracy during operational scoring. The competing hypothesis—that the training that occurs before the calibration itself is enough to recover skills that may have been lost over the gap in scoring—would be supported if there were no decrements in scoring accuracy for either calibration or validity samples.

To evaluate the two hypotheses, the data were analyzed using binomial linear mixed-effect models. All analyses were carried out in the R programming language and environment (R Development Core Team, 2008) using the lme4 software package (Bates, Maechler, & Dai, 2008). Table 4 shows the estimates (reported in log odds), standard errors, z-statistics,

Table 3 Means for Rater Days in Scoring, Days in Gap, and Accuracy Metrics for Calibration and Validity Scores

Variable and statistic	Mean	SE
Count of scoring days	67.21	2.48
Count of days in scoring gap	244.45	4.14
Average days in gap	7.23	0.47
Days of scoring consecutively	1.83	0.11
Proportion passed calibration on first attempt	0.90	0.00
Average seconds to score a calibration response	190.80	3.42
Difference between calibration scores and correct scores	0.01	0.00
Percentage exact on calibration	0.86	0.01
Percentage adjacent on calibration	0.13	0.01
Percentage discrepant on calibration	0.00	0.00
Average seconds to score a validity response	203.82	3.39
Difference between rater validity scores and correct scores	−0.10	0.00
Percentage exact on validity	0.67	0.00
Percentage adjacent on validity	0.32	0.00
Percentage discrepant on validity	0.01	0.00

Note. All means, except for the difference between calibration and correct scores, $p = .02$, are significantly different from zero, $p < .001$, a Bonferroni correction was used to correct for multiple comparisons.

Table 4 Estimates and z-Values of the Final Binomial Linear Mixed-Effects Models

Variable	B	SE	z	p
Exact on calibration samples				
(Intercept)	1.69	0.04	41.30	<0.001
Time spent on each calibration response	−0.27	0.009	−29.80	<0.001
Days since last scoring session	−0.001	0.0007	−2.01	0.044
Exact on validity samples				
(Intercept)	1.05	0.035	29.806	<0.001
Time spent on each calibration response	−0.11	0.007	−17.133	<0.001
Days since last scoring session	−0.002	0.0006	−3.058	0.002

Note. Model 5 for calibration; Model 11 for validity.

and p-values of the binomial linear mixed-effects models. Model comparisons were made by comparing models' Akaike information criteria (AICs).

Two separate sets of models were constructed to evaluate calibration and validity accuracy. In the first set of models, we focused on calibration (Models 1–6), and in the second set (Models 7–12), we focused on validity accuracy. In our analyses, we log-transformed seconds of time spent scoring each calibration (and validity) response and centered all of the three independent variables.

Calibration

The first two models we constructed investigated the effects of random factors so that we could determine which of these factors should be included in the full model predicting calibration accuracy as measured as scoring exact on the calibration sample. Model 1 predicted exact scores on calibration samples and included the rater identification number as a random factor. Subsequently, in Model 2, we added the item response identification number as a random factor. Model 2 was a better fit ($\Delta\text{AIC} = 32{,}006 > 2$). In Model 3, we also added the prompt identification number as a random factor. However, adding the prompt identification number did not improve the model ($\Delta\text{AIC} = 0.13 < 2$). Therefore, for the subsequent models, for predicting calibration, we only used the rater identification number and item response identification number as random factors.

After establishing suitable random factors, independent variables were added hierarchically to evaluate a complete model. The independent variables in which we were interested for predicting calibration accuracy were seconds of time spent scoring each calibration response, days since the last scoring session, and the number of consecutive days of scoring up to that day. In Model 4, we first added the seconds of time spent per calibration response and compared the model with Model 2 (see earlier; ΔAIC = 881 > 2, which indicated that Model 4 was a better fit). In Model 5, we also added the days since the last scoring session. Model 5 was a slightly better fit than Model 4 (ΔAIC = 2), and so the days since last scoring session was retained as a variable in our model. Subsequently, we added the number consecutive days of scoring in Model 6. However, adding consecutive days of scoring did not improve the model (ΔAIC = 0.29 < 2).

Results of the model fitting revealed that time spent scoring each calibration response and days since last scoring session were significant predictors of calibration accuracy. More time spent on each calibration response was related to raters scoring fewer responses as exact. There was also a negative relationship between days since last scoring session and calibration scoring accuracy, but the effect size was weak, and thus the results should be interpreted with caution.

The number of consecutive days of scoring did not significantly predict calibration accuracy.

Validity

As with calibration, the first two models we constructed investigated the effects of random factors so that we could determine which of these factors should be included in the full model predicting validity accuracy as measured as scoring exact on the validity sample. Model 7, which predicted exact scores on validity samples, included the rater identification number as a random factor. Subsequently, in Model 8, we added the response identification number as a random factor. Model 8 was a better fit (ΔAIC = 58,407 > 2). In Model 9, we also added the prompt identification number as a random factor. Adding the prompt identification number improved the model (ΔAIC = 137 > 2). Therefore, for the subsequent models, for predicting exact on validity samples, we used the rater identification number, response identification number, and prompt identification number as random factors.

In Model 10, we added the seconds of time spent scoring each validity item and compared with Model 9 (ΔAIC = 285 > 2, so Model 10 was a better fit). In Model 11, we also added days since last scoring session as an independent variable. Model 11 was better than Model 10 (ΔAIC = 7 > 2). Finally, we added consecutive days of scoring in Model 12. However, adding consecutive days of scoring did not improve the model (ΔAIC = −1.48 < 2).

Results of the model fitting demonstrated a similar pattern to that shown with calibration. More time spent scoring each validity response was related to raters scoring fewer validity samples as exact. As with calibration, there was also a negative relationship between days since last scoring session and validity scoring accuracy. Although the effect size was not large, it significantly predicted scoring exact. As with calibration, consecutive days of scoring did not significantly predict validity scoring accuracy.

Discussion

In our investigation of how the time between scoring sessions influences scoring accuracy, we found that the number of days of consecutive scoring did not impact calibration accuracy or validity sample accuracy. That is, the number of consecutive days of scoring was neither helpful nor hurtful to performance. The time spent scoring each response was a significant predictor of exact scoring for both validity and calibration scoring accuracy. More time spent scoring validity samples was related to a decrease in scoring accuracy for both calibration and validity samples.

A gap in days of scoring was also associated with a decrease in scoring exact for both calibration and validity samples. This pattern of results was in line with a decline in skills theory, which posited that both calibration and validity sample accuracy should decline as the scoring gap increases. The results did not align with the calibration-as-retraining theory, which predicted that even if skills did decline over a gap in scoring, there should be a decline neither in calibration nor in validity scoring accuracy, because the calibration session would serve as a retraining exercise.

However, the results from the calibration and validity models suggest that the number of days since the last scoring session matters more for validity accuracy than it does for calibration accuracy. One possible explanation is that during calibration and validity scoring, raters engaged in distinct cognitive processes, which could be differentially affected by gaps in scoring. Calibration tests typically require that raters meet a minimum performance standard before they are allowed to move on to operational scoring. A criterion test of this nature may involve different scoring strategies and/or attention directed toward distinct cues and information than might be employed during operation scoring. For example, during calibration, raters may be more attentive to the rubric and scoring guidelines than they are during operational scoring. Indeed, Joe et al. (2011) demonstrated that once raters in their study (both experienced and inexperienced) began scoring operationally, they rarely consulted the scoring rubric.

Furthermore, in our data, the time spent scoring calibration samples was significantly less than time spent on validity samples ($M = 190.80$ seconds, SE = 3.42 vs. $M = 203.82$ seconds, SE = 3.39, for calibration and validity scoring, respectively), suggesting that processing during the different scoring phases indeed is distinct. One possibility is that during operational scoring, raters are highly sensitive to the fact that they are scoring responses that have high stakes for the test takers. They may take more time to score each response than they spend scoring during calibration because they want to “get it right.” A second nonmutually exclusive possibility is that during calibration, raters are allowed to “misscore” a couple of samples and still pass calibration. A rater might proceed through calibration more quickly so that he or she can move on to actual operational scoring. Rater cognition remains an underexplored topic, however (e.g., Suto, Crisp, & Greatorex, 2008) and questions concerning the differences in cognition processes employed during calibration and validity scoring remain to be answered.

In summary, our findings demonstrate that the number of consecutive calibrations does not show a relationship to nonoperational and operational scoring metrics. In contrast, time spent per response and the days in the gap between scoring sessions are negatively correlated with the percentage of validity samples scored exact. These findings point to the possibility that calibrating consecutively over several days may not provide an advantage to scoring accuracy above a less frequent calibration schedule. However, care should be taken to ensure that the number of days in the gap is not extensive; otherwise, operational accuracy could suffer. The findings are suggestive of these possibilities, but further empirical work must focus on untangling calibration from operational scoring accuracy over time. Therefore, finding an optimal calibration and operational scoring schedule is crucial. Another point to consider is that if raters are indeed treating calibration and operational scoring distinctly, then the operational phase should be considered as an important aspect of quality control, as it is a process to which the raters are blind. Finally, these analyses were conducted using a GRE data set. Research is currently being conducted with other program data that represent, among other distinctions, different prompt types and number of scoring levels. This follow-up work will enable us to determine the extent to which the current findings generalize to a distinct scoring context.

The optimization of practice schedules to maximize learning and retention has been one of the central research domains in intelligent tutoring systems. Intelligent tutoring systems are computerized instructional systems that scaffold and support student learning outcomes. Successful implementation of these systems traces students’ knowledge and skills and also takes into account students’ forgetting rate to predict their future performance (Jastrzembski, Gluck, & Gunzelmann, 2006; Pavlik & Anderson, 2003, 2005, 2008; Pavlik, 2007; van Rijn, van Maanen, & van Woudenberg, 2009). It would be ideal to adapt a similar approach and design a personalized calibration and/or operational scoring scheduling system that considers an individual rater’s forgetting rate and level of expertise, to minimize nonoperational time and to maximize operational scoring accuracy.

The current sample used the GRE rating pool, which is fairly stable and consists of fairly experienced raters. Further studies employing an experimental design examining the impact of reducing calibration frequency on scoring accuracy should also be undertaken. One worthwhile avenue for follow-up research would be to replicate this study with a rater pool that has a higher turnover rate or is scoring for a newer assessment program. Findings from this research will facilitate development of best practices for CR scoring.

Complementary follow-up work is aimed at further investigating rater cognition during calibration. For example, why does the time spent on scoring differentially influence the success of operational and calibration scoring? How do practices that have been shown to mitigate skill loss, such as overlearning, conditions of retrieval, evaluation criteria, and retraining (Arthur et al., 1998; Farr, 1987; Healy et al., 1993; Jastrzembski et al., 2006; Schmidt & Bjork, 1992), influence retention of essay rating skills? Much research remains to be done to clearly understand how rater cognition influences scoring accuracy.

Note

[1] We do not report measures of adjacency, discrepancy, or bias in the current report.

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

Bates, D., Maechler, M., & Dai, B. (2008). *lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-28* [Computer software]. Retrieved from http://lme4.r-forge.r-project.org/

Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement Issues and Practice, 14*(4), 21–24. https://doi.org/10.1111/j.1745-3992.1995.tb00885.x

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

Coffman, W. E. (1971). On the reliability of ratings of essay examinations in English. *Research in the Teaching of English, 5*, 24–36.

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

Deane, P. (2011). *Writing assessment and cognition* (Research Report No. RR-11-14). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02250.x

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York, NY: Dover. (Original work published 1885)

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

Farr, M. J. (1987). *The long-term retention of knowledge and skills: A cognitive and instructional perspective* (Report No. IDA-M-205). Alexandria, VA: Institute for Defense Analyses. https://doi.org/10.1007/978-1-4612-1062-7

Finn, B., & Roth, A. (2017). *An interview study with rSAT raters on calibration and operational scoring practices*. Unpublished manuscript.

Guilford, J. P. (1936). *Psychometric methods*. New York, NY: McGraw-Hill

Healy, A. F., Clawson, D. M., McNamara, D. S., Marmie, W. R., Schneider, V. I., Rickard, T. C., ... Bourne, L. E., Jr. (1993). The long-term retention of knowledge and skills. *Psychology of Learning and Motivation, 30*, 135–164. https://doi.org/10.1016/S0079-7421(08)60296-0

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the Golden State Examination. *Journal of Educational Measurement, 38*, 121–145. https://doi.org/10.1111/j.1745-3984.2001.tb01119.x

Jastrzembski, T., Gluck, K., & Gunzelmann, G. (2006). Knowledge tracing and prediction of future trainee performance. In *Proceedings of the 2006 Interservice/Industry Training, Simulation, and Education Conference* (pp. 1498–1508). Orlando, FL: National Training Systems Association.

Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: A mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy, and Practice, 18*, 239–258. https://doi.org/10.1080/0969594X.2011.577408

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. A. (2010). Constructed-response scoring—Doing it right. *R&D Connections, 13*. Retrieved from https://www.ets.org/Media/Research/pdf/RD_Connections13.pdf

Melton, A. W. (1967). Repetition and retrieval from memory. *Science, 158*, 532. https://doi.org/10.1126/science.158.3800.532-b

Murre, J. M., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PLoS One, 10*(7). https://doi.org/10.1371/journal.pone.0120644

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

Pavlik, P. I., Jr. (2007). Understanding and applying the dynamics of test practice and study practice. *Instructional Science, 35*, 407–441. https://doi.org/10.1007/s11251-006-9013-2

Pavlik, P. I., Jr., & Anderson, J. R. (2003). An ACT-R model of the spacing effect. In *Proceedings of the 5th International Conference of Cognitive Modeling* (pp. 177–182). New York, NY: Springer.

Pavlik, P. I., Jr., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*, 559–586. https://doi.org/10.1207/s15516709cog0000_14

Pavlik, P. I., Jr., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied, 14*, 101–117. https://doi.org/10.1037/1076-898X.14.2.101

R Development Core Team. (2008). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Ricker-Pedley, K. L. (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

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413 – 428. https://doi.org/10.1037/0033-2909.88.2.413

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*, 207 – 217. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

Suto, I., Crisp, V., & Greatorex, J. (2008). Investigating the judgemental marking process: An overview of our recent research. *Research Matters, 5*, 6 – 9.

van Rijn, D. H., van Maanen, L., & van Woudenberg, M. (2009). *Passing the test: Improving learning gains by balancing spacing and testing effects*. Paper presented at the 9th International Conference of Cognitive Modeling, Manchester, England.

Wigglesworth, G. (1994). Patterns of rater behaviour in the assessment of an oral interaction test. *Australian Review of Applied Linguistics, 17*(2), 77 – 103. https://doi.org/10.1075/aral.17.2.04wig

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R&D Connections, 21*, 2 – 13.

Suggested citation:

Finn, B., Wendler, C., Ricker-Pedley, K. L., & 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

Action Editor: Anastassia Loukina

Reviewers: Lawrence Davis and Jung Aa Moon