

# A Comparison of Training & Scoring in Distributed & Regional Contexts—Writing

Edward W. Wolfe

Staci Matthews

Daisy Vickers

Pearson

July 2009

**PEARSON**

The Pearson logo consists of the word "PEARSON" in a bold, blue, sans-serif font. Below the text is a yellow swoosh that starts under the "P", goes under the "A", and ends under the "N".

*Using assessment  
and research to  
promote learning*

## **Abstract**

This study examined the influence of rater training and scoring context on the following outcomes: (a) training time, (b) scoring time, (c) qualifying rate, (d) quality of ratings, and (e) rater perceptions. 120 raters participated in the study and experience one of three training/scoring contexts: (a) online training in a distributed scoring context (OD), (b) online training in a regional scoring context (OR), and (c) stand-up training in a regional context (SR). After training, raters assigned scores to qualification sets, scored 400 student essays, and responded to a questionnaire that measured their perceptions of the effectiveness of and satisfaction with the training and scoring process, materials, and staff. The results suggest that the only clear difference on the outcomes for these three groups of raters concerned training time—online training was considerably faster. There were no clear differences between groups concerning qualification rate, rating quality, or rater perceptions.

Note that a companion report is also available, which reports the results of a similar study conducted with raters of reading assessment constructed-response items based on a design that is similar to the one summarized in this report and producing comparable results.

## **A Comparison of Training & Scoring in Distributed & Regional Contexts—Writing**

Human scoring of products created based on constructed response assessment items may take place in several contexts. However, little research has been conducted that has focused on how features of the training and scoring context may impact the quality of scores that are assigned by human raters. This report summarizes the results of a study that compares the scoring of writing assessment performances under three conditions: (a) rater training that is conducted online followed by scoring that occurs through a computer interface at remote locations (referred to here as an *online distributed* scoring context), (b) rater training that is conducted online followed by scoring that occurs through a computer interface, both of which take place at a regional scoring center (referred to here as an *online regional* scoring context), and (c) face-to-face training followed by scoring that occurs through a computer interface, both of which take place in a regional scoring center (referred to here as a *stand-up regional* context).

### **METHOD**

Data for this study were collected from 40 raters under each of these three conditions ( $n = 120$ ), each rater participating in only one of the three scoring contexts as part of a special study of these scoring contexts. Raters participated in training activities, assigned scores to qualifying sets, and scored a common set of 400 student essays through an online distribution system. Performance on all of these scoring tasks, in addition to the amount of time required to complete training and scoring, was documented. Raters also responded to a questionnaire designed to document demographic, educational, and professional characteristics; as well as rating scales designed to document their perceptions of the effectiveness of and their satisfaction with the

training and scoring materials, procedures, and personnel. Scoring supervisors also documented the number and nature of requests for assistance that were made by raters.

## Participants

The Human Resource Team for Pearson's Performance Scoring Center secured participation in the study from three groups of raters, selected to be comparable in several demographic (gender, age, ethnicity), educational (undergraduate major and highest level attained), and professional experience (scoring and teaching experience) variables. The participants had not previously scored essays for the operational project from which papers were selected for use in the current study. Participants were paid in lump sum for completing the training and scoring. The pay rates were equal regardless of the group into which a rater was placed.

Raters responded to a questionnaire that documented several demographic (gender, age, ethnicity), educational (undergraduate major and highest level attained), and professional experience (scoring and teaching experience) variables. Because raters could not be randomly assigned to conditions (due to geographic restrictions), it was important to verify that the three groups are comparable with respect to relevant demographic characteristics in order to warrant comparison of group performance statistics.

**Table 1** indicates that the three scoring context groups were fairly comparable with respect to demographics, educational attainment, and professional experiences. With respect to demographic characteristics, participants in the online distributed group were slightly more likely to be female and under the age of 55 while participants in the other two groups were more likely to be white. However, these differences were not statistically significant:  $\chi^2_{(2) \text{ Gender}} = 1.27, p = .52$ ;  $\chi^2_{(4) \text{ Age}} = 7.66, p = .09$ ; and  $\chi^2_{(8) \text{ Ethnicity}} = 13.73, p = .05$ . With respect to education, the

online distributed scorers were more likely to have non-response data for undergraduate major and to have attained a graduate degree than the other two groups. The difference between the reported undergraduate major for raters in the three scoring context groups was statistically significant [ $\chi^2_{(8)} = 22.07, p = .006$ ] while the difference for graduate degree was not [ $\chi^2_{(4)} = 7.61, p = .13$ ]. Finally, with respect to teaching experience, the online distributed scorers were more likely to have previously participated in four or more scoring projects, and the online regional scorers were less likely to have secured a teaching certification. However, neither of these differences is statistically significant:  $\chi^2_{(4)} \text{ Scoring Experience} = 1.88, p = .78$ ;  $\chi^2_{(2)} \text{ Teaching Certificate} = 0.40, p = .88$ , respectively.

**Table 1: Demographic, Education, and Experience by Scoring Context**

Variable	Level	Online Distributed	Online Regional	Stand-up Regional
Gender				
	Female	58% (23)	45% (18)	48% (19)
	Male	43% (17)	55% (22)	52% (21)
Age				
	Under 30	10% (4)	8% (3)	5% (2)
	30 to 55	58% (23)	33% (13)	33% (13)
	55 or older	33% (13)	60% (24)	58% (23)
Ethnicity				
	Asian	6% (2)	3% (1)	0% (0)
	Black	17% (6)	3% (1)	8% (3)
	Hispanic	6% (2)	3% (1)	5% (2)
	White	71% (25)	87% (35)	88% (35)
	No Response	(5)	(2)	(0)
Undergraduate				
Major	Business	28% (9)	35% (14)	38% (15)
	Humanities/Liberal Arts	63% (20)	53% (21)	46% (23)
	Sciences	9% (3)	13% (5)	5% (2)
	No Response	(8)	(0)	(0)
Highest Education				
Level Attained	Bachelor	55% (22)	68% (27)	75% (30)
	Masters	35% (14)	13% (13)	23% (9)
	Doctoral	10% (4)	0% (0)	3% (1)
Scoring				
Experience	New	5% (2)	10% (4)	13% (5)
	1 to 3 projects	23% (9)	28% (11)	25% (10)
	4 or more projects	73% (29)	63% (29)	63% (25)
Teaching				
Certification	Yes	23% (9)	18% (7)	23% (9)
	No	78% (31)	83% (33)	78% (31)

## **Materials & Procedures**

The training materials for this project (online training modules and anchor, practice, and qualification responses) were originally developed for the stand-up training used in the operational scoring project from which responses for this study were sampled. The scoring rubric upon which scores were based was a four-point, focused holistic rubric. Members of the range-finding committee assigned consensus scores to responses which were compiled into two sets of 10 practice papers (completed by raters during training) and three sets of 10 qualifying papers (scored by raters at the conclusion of training but prior to scoring). A senior content specialist, familiar with both online and stand up training reviewed the materials and made adjustments for online training. A full range of scores was represented in each group of training materials. All scoring directors completed the online training modules and online practice and qualification sets. With the exception of the fact that those participating in online training viewed images of the original response while those participating in stand-up training viewed photocopies of the original response, the training materials were the same for online and stand-up training. The stand-up trainer used standardized annotations written for each response to explain the rationale for the consensus scores in order to minimize the introduction of additional concepts or verbiage (beyond what was presented in the online training) in the stand-up training group.

For the scoring component of the study, 600 responses were pulled at random from the operational assessment for each of the items, and each response was scored independently by at least three of the scoring directors. The scoring directors then worked together to choose the 400 responses raters in the study would score, with instructions to choose a variety of responses spanning the score point scale, eliminating blank or off-topic responses and responses that were

less representative of the response types most seen in scoring. The scoring directors also chose a set of calibration (retraining) papers.

In the online training that was used with distributed raters and regional raters, the raters were expected to complete the training at their individual paces. For the stand-up training in the regional site, the raters were led through a training session from the front of the room with paper training materials. Members of the stand-up group progressed through training as a group at the same pace. At the regional site, raters could ask questions about the responses, either online or by going directly to a supervisor, and either the scoring director or a scoring supervisor would answer the question. For the distributed raters, scoring directors and scoring supervisors would respond to questions online or by phone. Supervisory staff in all three groups documented questions and interventions.

## **Measures**

In addition to the demographic questionnaire, data were collected relating to rater performance on several tasks, the amount of time required to complete training and scoring, rater perceptions of the effectiveness of and their satisfaction with the training and scoring context they experienced, and the number and nature of requests for assistance that were made by raters during the training and scoring process.

***Time:*** Scoring and training time were defined as the number of hours required to complete training for the project and to complete the scoring. The number of hours spent reviewing training materials and responding to qualifying sets was designated as the amount of ***training time***. For online distributed and online regional raters, this time was recorded by the online scoring system used to distribute training materials to the raters and to record their performance on the qualifying sets. For stand-up regional raters, the time was constant for all

raters because they participated in a group training setting and responded to qualifying sets during a common time frame. **Scoring time**, measured in hours, was automatically recorded by the online scoring system used to document the scores all raters assigned to each essay.

**Rater Performance:** The accuracy and agreement rates for raters in each group were measured in several ways. Performance on qualification sets was measured as the percentage of assigned scores that matched those assigned by rater trainers (**qualifying agreement**) as well as whether performance across three qualifying sets would have allowed raters to “qualify” for a scoring project (**qualification rate**), which is accomplished when a rater attains an agreement rate of 70% or better on either one (i.e., a typical qualifying standard) or two (i.e., a high qualifying standard) of the three qualifying sets.

Reliability and validity were measured in four ways in this study. First, **inter-rater reliability** was defined as the correlation between the scores assigned by a particular rater and the average score assigned by all other raters in the project to the 400 essays. This index indicates whether a particular rater rank ordered examinee responses in a manner that is consistent with the typical rank ordering of those examinees across the remaining raters in the study. Second, the **validity coefficient** was defined as the correlation between the scores assigned by a particular rater to the 400 essays and the consensus score assigned by scoring project leaders to those essays. Third, the **validity agreement index** was defined as the percentage of exact agreement between the scores assigned by raters to the 400 essays papers and the consensus scores assigned by project leaders. Fourth, **backreading agreement** was defined as the percentage of agreement raters had with scoring supervisors (project leaders who were not part of the process of selecting and assigning consensus scores to the papers used in training, and qualification) who read and rescored a small and variable proportion of the essays scored by each rater.

***Rater Perceptions:*** Rater perception of the effectiveness of training and scoring procedures and the level of rater satisfaction with their training and scoring experiences were measured with two fifteen-item questionnaires, each requesting that raters rate on a three-point scale ranging from 0 = “not very effective/satisfied” to 1 = “moderately effective/satisfied” to 2 = “very effective/satisfied” to various features of the scoring and training context (e.g., training procedures & materials, personnel, qualifying process, scoring process, scoring materials, etc.). Coefficient alpha for the effectiveness and satisfaction scales equals  $\alpha = .95$  and  $\alpha = .96$ , respectively.

***Requests for Assistance:*** Scoring supervisors kept logs to record the number and nature of requests for assistance made by raters. In the online distributed rater context, raters could seek assistance via telephone. Online regional raters could request assistance via e-mail or in person, while stand-up regional raters could request assistance in person. Request for assistance logs were perused for the sake of collecting anecdotal accounts of differences in the frequency and nature of such requests for raters in each scoring context.

## **Analyses**

For all outcome variables, scoring context was treated as an independent variable, and the analyses focused on determining whether groups differed on each outcome variable. When possible, planned comparisons were conducted, comparing the ***online distributed*** and the ***online regional*** raters’ performances to the performance of the ***stand-up regional*** reference group. All analyses adopted a Type I error rate of .05. When possible, effect size indices were computed for statistically significant outcomes, and these indices include  $\delta$  for each t-test,  $\eta^2$  for each Analysis of Variance (ANOVA), and conditional percentages for logistic regressions.

**Time:** A one-sample t-test was conducted to determine whether the *online distributed* and the *online regional* training/scoring context groups' number of training hours differed from the constant number of hours spent in training by the *stand-up regional* group. An ANOVA was conducted to determine whether the training/scoring context groups differed with respect to the number of hours spent scoring.

**Rater Performance:** An ANOVA was conducted to compare the performance of the training/scoring context groups on an arcsine transformation of qualifying rate agreement (measured as a percentage). A logistic regression was conducted to determine whether the training/scoring context groups differed with respect to qualification rate (measured as a dichotomous outcome—qualified versus did not qualify under the two scenarios of the standard one-of-three and the more demanding two-of-three sets with 70% or better agreement with scoring project leaders). T-tests were conducted to evaluate training/scoring context group differences on Fisher transformations of the inter-rater reliability and validity coefficients, and an ANOVA was conducted to determine whether the training/scoring context groups differed on an arcsine transformation of validity agreement index (measured as a percentage). An ANOVA was conducted to determine whether the training/scoring context groups differed with respect to an arcsine transformation of backreading agreement (measured as a percentage).

**Rater Perceptions:** An ANOVA was conducted to determine whether the training/scoring context groups differed with respect to measures of perceived effectiveness and rater satisfaction with training and scoring procedures, materials, and personnel. Because of a data coding anomaly, raters in the online distributed group represented a mixed group of raters including raters who participated in a companion study focusing on reading in addition to those participating in this (writing) study.

*Requests for Assistance*: Frequencies of assistance requests were summarized with no inferential methods being applied to these data due to concerns about the completeness of these data. Logs were perused, and entries were coded according to whether raters' requests for assistance concerned writing *content* and application of the scoring guidelines, questions or problems navigating the online computer *interface*, or questions about the *logistics* of training and scoring.

## RESULTS

### Training & Scoring Time

**Table 2** summarizes the number of training and scoring hours for each training/scoring context group. The number of hours of training for the stand-up regional group was considerably greater than that required for the two online training conditions, and these differences are statistically significant with large effect sizes, according to Cohen's guidelines (1988). Generally, stand-up training took about three times longer than the online training. With respect to scoring time, neither of the comparisons between the online distributed and online regional raters versus the stand-up regional raters was statistically significant, although the online distributed raters took slightly less time to finish scoring the 400 essays.

**Table 2: Scoring and Training Time by Group**

Variable	Statistics	OD	OR	SR
Training Time	Mean	3.40	4.75	12.00
	SD	1.40	1.37	NA
	$t_{vs. SR}$	38.84	33.52	
	$\delta$	6.14	5.30	
Scoring Time	Mean	14.95	18.75	17.41
	SD	7.98	5.29	4.67
	$F_{vs. SR}$	3.19	0.95	
	$p$	.08	.33	

Note: OD = Online Distributed, OR = Online Regional, and SR = Stand-up Regional.  $n = 40$  for all groups. NA = Not Applicable. For the t-tests,  $df = 39$  and  $p$  values are  $< .0001$ . For the F tests,  $df = (1, 119)$ .

### Qualifying Set Performance

**Table 3** presents the average percent of exact agreement between raters in each group and the consensus scores scoring project leaders assigned to the qualification papers as well as the qualification rates for each group. In terms of qualifying agreement rates, online distributed and stand-up regional raters performed somewhat better than online regional raters. The difference between online regional and stand-up regional was statistically significant with a moderately large effect size. Similar results were obtained for qualification rates. For the typical qualifying standard, which requires raters to obtain 70% or better agreement on one or more (out of three) qualifying sets, the average qualifying rates for the three training/scoring context groups are equal, so no inferential statistical test was conducted. For the more demanding qualifying standard of at least two of three qualifying sets at 70% or better agreement, although the online distributed raters attained the highest qualification rate, that rate was not greater than that of the stand-up regional group by a statistically significant degree. On the other hand, the qualifying

rate of the online regional group was less than that of the stand-up regional group by a statistically significant amount.

**Table 3: Qualifying Set Performance by Group**

Variable	Statistics	OD	OR	SR
Qualifying Set Agreement	Mean	79%	71%	79%
	SD	10.07	10.97	13.57
	$F_{vs. SR}$	0.10	10.09	
	$p$	.75	.001	
	$\eta^2$	--	.08	
Qualifying Rate	Mean	98%	98%	98%
	SD	15.81	15.81	15.81
1 of 3 sets	Mean	90%	70%	82%
	SD	30.38	46.41	38.48
2 of 3 sets	$\chi^2_{vs. SR}$	2.84	4.55	
	$p$	.09	.03	

Note: OD = Online Distributed, OR = Online Regional, and SR = Stand-up Regional.  $F$  tests were conducted on an arcsine transformation of qualifying set agreement.  $df = (1,357)$  for all  $F$  tests.  $df = (1)$  for all  $\chi^2$  tests.

### Reliability & Validity Performance

**Table 4** displays the score quality indices (i.e., the inter-rater reliability correlation, the validity correlation coefficient, and the validity percentage of agreement index) for each training/scoring context group. Overall, the online distributed group exhibited slightly better performance than the two regional training/scoring groups. However, none of the observed differences were statistically significant according to the  $z$  test conducted on the Fisher transformations of the inter-rater reliability and validity coefficients or the ANOVA conducted on the arcsine transformation of the validity agreement index.

**Table 4: Reliability and Validity Performance by Group**

Variable	Statistics	OD	OR	SR
Interrater Reliability	Mean	.81	.78	.77
	SD	.06	.05	.07
	$z_{vs. SR}$	0.46	0.11	
	$p$	.32	.46	
Validity Coefficient	Mean	.72	.70	.69
	SD	.05	.06	.07
	$z_{vs. SR}$	0.26	0.08	
	$p$	.40	.47	
Validity Agreement Index	Mean	57%	58%	57%
	SD	6.99	5.91	6.94
	$F_{vs. SR}$	0.00	0.15	
	$p$	0.96	0.70	

Note: OD = Online Distributed, OR = Online Regional, and SR = Stand-up Regional.  $z$  tests were conducted on a Fisher transformation of interrater reliability and validity coefficients.  $F$  tests were conducted on an arcsine transformation of backreading agreement.  $df = (1, 119)$  for the  $F$  tests.

### Backreading Agreement

**Table 5** displays descriptive statistics for backreading rate and backreading agreement by training/scoring context group. Backreading rate was considerably greater in the online distributed group than in the two regional groups. Anecdotal reports suggest that this was because the online distributed scoring leaders spent less time interacting with readers, leaving more time available to conduct backreading. The online regional group exhibited the highest level of backreading agreement among the three groups. The difference between the backreading agreement rates of online distributed and stand-up regional raters' backreading agreement rates was not statistically significant. However, the backreading agreement rate of the online regional raters was higher than those of the stand-up regional by a statistically significant amount.

**Table 5: Backreading Agreement by Group**

Variable	Statistics	OD	OR	SR
Backreading Rate	Mean	47.15	22.03	24.15
	SD	19.69	10.51	10.43
Backreading Agreement	Mean	73.77	83.28	74.86
	SD	12.91	11.57	14.27
	$F_{vs. SR}$	0.34	8.75	
	$p$	0.56	0.004	
	$\eta^2$	--	.07	

Note: OD = Online Distributed, OR = Online Regional, and SR = Stand-up Regional.

$df = (1, 119)$  for the  $F$  tests.

### Perception of Training and Scoring

**Table 6** displays the average score on the two rater perception scales for the three training/scoring context group. On both scales, measures for the online distributed and stand-up regional groups were slightly higher than those of the online regional group. However, none of these differences was statistically significant.

**Table 6: Training and Scoring Perception Measures by Group**

Variable	Statistics	OD	OR	SR
Effectiveness	Mean	1.61	1.54	1.63
	SD	0.41	0.45	0.43
	$F_{vs. SR}$	0.02	0.37	
	$p$	0.89	0.55	
Satisfaction	Mean	1.60	1.53	1.64
	SD	0.43	0.44	0.50
	$F_{vs. SR}$	0.06	0.50	
	$p$	0.81	0.48	

Note: OD = Online Distributed, OR = Online Regional, and SR = Stand-up Regional.  $n_{OD} = 49$ ,  $n_{OR} = 22$ ,  $n_{SR} = 12$ . For each F test,  $df = (1, 82)$ .

### Requests for Assistance

**Table 7** displays the counts (and percentages) for each of three types of requests for assistance that were recorded in scoring supervisors' logs. Although inferential statistical tests were not conducted on these data due to concerns about the completeness of the supervisor logs, the figures suggest that raters in the stand-up regional condition were more likely to initiate requests for assistance, and those requests were more likely to focus on writing content than was the case for the online rater groups. On the other hand, the online distributed raters seemed more likely to request assistance for issues relating to training and scoring logistics while online regional raters seemed more likely to request assistance for issues relating to the online training and scoring computer interface.

**Table 7: Request for Assistance Focus by Group**

<b>Focus</b>	<b>OD</b>	<b>OR</b>	<b>SR</b>
Content	4 (24%)	5 (19%)	31 (48%)
Interface	2 (12%)	16 (59%)	9 (14%)
Logistics	11 (65%)	6 (22%)	25 (38%)
<b>Total</b>	<b>17</b>	<b>27</b>	<b>65</b>
<b>(Mean)</b>	<b>(0.43)</b>	<b>(0.68)</b>	<b>(1.63)</b>

Note: OD = Online Distributed, OR = Online Regional, and SR = Stand-up Regional.

## **DISCUSSION & CONCLUSIONS**

These results suggest several points concerning the nature of online and stand-up training and the distributed and regional scoring contexts. *First, training time may be shorter when delivered online, but scoring time is not greatly impacted by scoring context.* With respect to training time, it seems that stand-up training may take up to three times longer to complete than online training (about 4 hours for online training versus about 12 hours for stand-up training). The human interactions required for stand-up training are likely the cause of this difference, given that the differences exist between the stand-up regional group and both of the online groups of raters. Trainers in the stand-up context likely spend time introducing themselves, answering questions from individuals, and manipulating materials. These are tasks that are not required in an online training system. In addition, because of the nature of group training the speed of all raters is slowed to accommodate the slowest of the group. It is also noteworthy that the materials were originally developed for stand-up training and that the process of adapting those materials for online delivery did not take advantage of potential enhancements that might

be available through the use of technology to deliver the training. Therefore, it is possible that the observed differences in this study are an underestimate of the increased efficiency of training that may be realized through online training.

Concerning scoring time, although there was no statistically significant difference, the amount of time required for online distributed scoring was slightly less than for the two regional contexts—about 15 hours for online distributed and about 18 for the two regional context groups. Again, it may be that the context-bound human interactions of being “on site” is the cause of this slight increase in scoring time. It is also worth noting that, in this study, we did not account for calendar time required to complete the project. That is, although we can determine the number of calendar days required to complete the entire scoring project for raters in a regional context (i.e., add training and scoring time and divide by the number of hours in a work day—that is the number of days from the beginning of the project until each rater completed the work), the number of hours spent “on task” for raters in the distributed context could either underestimate or overestimate the number of calendar days that would be required for those raters to complete the scoring project. For example, if those raters could log only a minimal amount of time per day, it may take a more calendar days for those raters to complete the project. On the other hand, if more of raters in the distributed context were available due to the wider geographic based upon from which raters could be recruited, then it may be that those raters could complete the project more quickly due to their additional hours being worked by the additional raters.

***Second, scoring context may not influence the immediate performance of raters following training.*** Our data indicate that agreement rates on qualifying sets were equivalent for online distributed and stand-up regional raters (79%), and both of these groups performed slightly better than raters in the online regional setting (71%). When these numbers were

translated into typical qualification standards (i.e., 70% or better agreement on at least one of three qualifying sets), there were no differences in the performance of raters in the three groups—about 98% of the raters in each group achieved that standard. On the other hand, raters in the online distributed and stand-up regional contexts were more likely to achieve the higher qualifying standard (i.e., 70% or better agreement on at least two of three qualifying sets) than were the online regional raters. Because backreading agreement was lower for the online regional group and because rater perceptions of the training/scoring materials, procedures, and staff was less positive, we speculate that these results represent an idiosyncrasy associated with the way the scoring leaders in the online regional interacted with their raters. Alternatively, it may be that the coupling of human interactions with the delivery of training materials in an online context led to some confusion in the online regional training/scoring context.

***Third, the training/scoring context does not seem to influence the quality of ratings.*** In this study, raters in the online distributed context were better able to agree with one another (i.e., higher interrater reliability) and were better able to agree with scoring leaders (i.e., higher validity coefficients). However, none of the observed differences were statistically significant. Hence, it seems that the online training context is considerably more efficient than the stand-up context. Not only did raters in the online groups complete training more quickly, but the scores that they assigned were of equal, if not better, quality than those assigned by raters who experienced stand-up training.

***Fourth, training/scoring context may not influence rater perceptions of the training and scoring process.*** In this study, raters who experienced the online regional training/scoring context had less positive views of the effectiveness of and lower satisfaction with the training

and scoring materials, procedures, and personnel when compared to the other two training/scoring groups. However, the observed differences were not statistically significant.

*Finally, scoring context may impact other variables, such as backreading rates and the number and nature of requests for assistance that raters make.* Analysis of backreading rates indicates that online distributed scoring project leaders spent more time engaging in backreading when the raters were in a distributed context. It is likely that the nature of regional scoring requires scoring leaders to spend more time engaging in interactions with raters (e.g., requests for assistance) and handling logistical issues than is the case for distributed scoring. In fact, our analyses of the amount and nature of assistance that is requested from scoring supervisors by raters supports this notion. Raters in the stand-up regional training/scoring context requested such assistance more frequently than raters in the online contexts, and their requests were more likely to focus on writing content questions rather than online computer interface issues or logistics of the scoring project—recall that this group did not perform as well in terms of validity scores as those trained in an online context. Raters in the online distributed group made more requests for assistance for issues relating to logistics, while those in the online regional group made more requests relating to the online scoring interface.

It is important to interpret these differences in light of the fact that these data come from intact groups—we were unable to randomize training/scoring context in our study. Hence, one should keep in mind that a potential alternative explanation of the observed differences is due to the fact that *the three groups of raters differed slightly in terms of demographic, educational, and professional experience variables.* However, we should emphasize that the only statistically significant difference between groups concerned undergraduate major (those in the online distributed group were more likely to have non-response data on this variable). Regardless, the

online distributed group did differ slightly from the two groups of regional raters in potentially important ways. Specifically, online distributed raters were more likely to be female (58% versus about 47% for the two regional groups), be between the ages of 30 and 55 (58% versus about 33% for the two regional groups), be black (15% versus about 3% for the two regional groups), have attained a graduate degree (45% versus about 20% for the two regional groups), and have scored in four or more previous scoring projects (73% versus 63%).

As mentioned above, it is possible that any group differences in rater performance that we observed may be a result of rater experiences associated with existing demographic, education, and professional experience differences. For example, one might hypothesize that raters with graduate degrees or more scoring experience, as is true for the online distributed group, might be better prepared for the scoring task. However, there is little published research that seeks to determine the relationship between rater characteristics and performance on scoring tasks such as the ones performed by raters in this study. Research relating to human-computer interactions indicates that younger people tend to show higher levels of facility with computer tasks than do older people (Colley & Comber, 2003). Conversely, that same body of research has indicated that blacks and more females may underperform on computer-based tasks (Cooper, 2006; Gallagher, Bridgeman, & Cahalan, 2002; van Braak & Kavadias, 2005). Hence, it is unclear whether the existing differences between groups can explain the observed performance differences because there is evidence that would both support and conflict with the notion that the online distributed group was advantaged by its composition.

However, even if existing group differences could explain the trends we discovered in our study, we believe that the differences that we observed in terms of demographic, education, and professional experience may reflect differences in the availability of raters between regional

and distributed contexts. That is, because of the restrictions associated with these contexts, it may be that the populations that are drawn upon for online versus regional scoring projects are different to begin with and that the observed rater performance differences reflect differences in the capabilities of those populations. Hence, any observed differences between the performance of raters in distributed and regional contexts observed in our study will likely manifest themselves in any operational setting due to the fact that the populations of available raters will be different for those two contexts.

At this time, it seems reasonable to conclude that online rater training, as implemented in the system employed in this study, is more efficient than the stand-up training employed, and that the online rater training system is at least as effective, if not more effective, than stand-up training. It is worth noting that rater performance and productivity was comparable between online and stand-up training, and may have been superior for those trained online, even though raters in all three contexts reported comparable levels of perceived effectiveness of and satisfaction with the training and scoring procedures and resources.

## References

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Colley, A., & Comber, C. (2003). Age and gender differences in computer use and attitudes among secondary school students: what has changed? *Educational Research, 45*, 155-165.
- Cooper, J. (2006). The digital divide: The special case of gender. *Journal of Computer Assisted Learning, 22*, 320-334.
- Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial/ethnic and gender groups. *Journal of Educational Measurement, 39*, 133-147.
- van Braak, J., & Kavadias, D. (2005). The influence of social-demographic determinants on secondary school children's computer use, experience, beliefs and competence. *Technology, Pedagogy and Education, 14*, 43-60.