

# ARRSO

A Comparison of Training & Scoring in  
Distributed & Regional Contexts—Reading

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 experienced 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 responses on each of three reading prompts, 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 a writing assessment 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—Reading**

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 reading 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 ( $n = 120$ ), with 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 responses on each of three constructed response reading items 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 for the operational project from which student responses were selected for use in the current study. This was also true for scoring supervisors and for scoring directors who assigned consensus scores to the papers in the qualifying sets and all papers scored by scorers in the project. Participants were paid a 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 and white, when compared to participants in the other two

groups. However, these differences were not statistically significant:  $\chi^2_{(2) \text{ Gender}} = 1.88, p = .42$ ;  $\chi^2_{(4) \text{ Age}} = 6.50, p = .15$ ; and  $\chi^2_{(6) \text{ Ethnicity}} = 6.00, p = .31$ . 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. However, neither of these differences was statistically significant [ $\chi^2_{(8) \text{ UG Major}} = 3.48, p = .48$  and  $\chi^2_{(4) \text{ Graduate Degree}} = 2.50, p = .67$ ]. Finally, with respect to professional experience, the stand-up regional scorers were more likely to have previously participated in four or more scoring projects, and the online scorers were more likely to have secured a teaching certification. Again, neither of these differences is statistically significant:  $\chi^2_{(4) \text{ Scoring Experience}} = 2.69, p = .61$ ;  $\chi^2_{(2) \text{ Teaching Certificate}} = 0.88, p = .74$ , respectively.

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

Variable	Level	Online Distributed	Online Regional	Stand-up Regional
Gender				
	Female	74% (29)	65% (26)	60% (24)
	Male	26% (10)	35% (14)	40% (16)
	No Response	(1)	(0)	(0)
Age				
	Under 30	15% (6)	13% (5)	15% (6)
	30 to 55	62% (24)	43% (17)	38% (15)
	55 or older	23% (9)	45% (18)	48% (19)
	No Response	(1)	(0)	(0)
Ethnicity				
	Asian	10% (2)	3% (1)	8% (3)
	Black	15% (3)	38% (15)	28% (11)
	Hispanic	5% (1)	3% (1)	0% (0)
	White	70% (14)	58% (23)	65% (26)
	No Response	(20)	(0)	(0)
Undergraduate				
Major	Business	6% (2)	0% (0)	5% (2)
	Humanities/Liberal Arts	84% (26)	85% (34)	88% (35)
	Sciences	10% (3)	15% (6)	8% (3)
	No Response	(9)	(0)	(0)
Highest Education				
Level Attained	Bachelor	60% (24)	75% (30)	70% (28)
	Masters	33% (13)	23% (9)	25% (10)
	Doctoral	8% (3)	3% (1)	5% (2)
Scoring				
Experience	New	10% (4)	13% (5)	13% (5)
	1 to 3 projects	33% (13)	28% (11)	18% (7)
	4 or more projects	56% (22)	60% (24)	70% (28)
	No Response	(1)	(0)	(0)
Teaching				
Certification	Yes	15% (6)	13% (5)	20% (8)
	No	85% (34)	88% (35)	80% (32)

## **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. Scoring directors 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 distribution system used to document the scores all raters assigned to each student response.

**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, in this project, required a rater to attain 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 student responses for each reading prompt. 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 responses within a prompt and the consensus scores assigned by scoring project leaders to those student responses. Third, the **validity agreement index** was defined as the percentage of exact agreement between the scores assigned by raters to the 400 responses within a prompt 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 student responses 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  $\alpha = .91$  and  $\alpha = .93$ , respectively.

## **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 (individually) to the performance of the *stand-up regional* reference group. Also where possible, analyses employed a repeated measures (i.e., repeated observations of a rater across the three reading prompts) Analysis of Variance (ANOVA) in which item-by-group interactions were first examined, and group main effects were interpreted when those interactions were not statistically significant. Hence, where possible, our results summarize main effects for context group, collapsing across items. All analyses adopted a Type I error rate of .05. Effect size indices were computed for statistically significant outcomes, and these indices include  $\delta$  for t-tests,  $\eta^2$  for ANOVAs, and conditional percentages for logistic regressions.

**Time:** One-sample t-tests were conducted for each item to determine whether the *online distributed* and the *online regional* training/scoring context groups' number of training hours differed from the constant value of number of hours spent in training by the *stand-up regional* group. A repeated measures ANOVA was conducted to determine whether the training/scoring context groups differed on the number of hours spent scoring across reading prompts.

**Rater Performance:** A repeated measures ANOVA was conducted to determine whether the training/scoring context groups differed with respect to qualifying agreement (measured as a percentage). Due to data sparseness (i.e., a singular covariance matrix), inferential statistics could not be applied to compare the training/scoring context groups 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), so only descriptive statistics are presented for these variables. Because reliability and validity coefficients differed very little across items, t-tests were conducted on Fisher transformations of the averaged reliability and validity coefficients (averaged across items within each group) to evaluate training/scoring context group differences. A repeated measures ANOVA was conducted to determine whether the training/scoring context groups differed with respect to validity agreement (measured as a percentage). A repeated measures ANOVA was also computed to determine whether the training/scoring context groups differed with respect to backreading agreement (measured as a percentage).

**Rater Perceptions:** T-tests were 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 writing in addition to those participating in this (reading) study.

## RESULTS

### Training & Scoring Time

**Table 2** summarizes the number of training and scoring hours for each group. For training time, data are summarized separately for each item because one-sample t-tests were required for these data due to the fact that all standup regional raters had the same number of training hours for each item. Generally, the greatest amount of training time was recorded for the first item (29), and the number of hours required was less for the two online groups while it remained relatively high for the stand-up group. The differences between training times for the online distributed raters and for the online regional raters when compared to the fixed number of hours required for the stand-up regional group are all statistically significant, and the effect sizes ( $\delta$ ) are large, according to Cohen's guidelines (1988).

With respect to scoring time, data are reported for each reading prompt because the group-by-item interaction is statistically significant with a large effect size [ $F_{(4,234)} = 8.37, p < .0001, \eta^2 = .12$ ]. Generally, the amount of time required to score the reading prompts was relatively high for the online regional group for the first reading prompt (29), and the three groups exhibited scoring times that were more similar for the second and third prompts (30 and 31). **Table 2** shows that the stand-up regional group completed the first set of student responses in about 2 hours less time than the online regional group. The remaining comparisons were not statistically significant.

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

Variable	Prompt	Statistic	OD	OR	SR
Training Time	29	Mean	2.22	3.24	5.83
		SD	0.86	1.33	NA
		$t_{vs. SR}$	7.22	7.89	
		$\delta$	2.83	2.01	
	30	Mean	1.16	1.68	4.32
		SD	0.58	0.98	NA
		$t_{vs. SR}$	8.49	8.43	
		$\delta$	1.82	2.10	
	31	Mean	1.31	1.81	5.75
		SD	0.61	0.87	NA
		$t_{vs. SR}$	8.61	8.73	
		$\delta$	1.67	1.31	
Scoring Time	29	Mean	5.13	6.05	4.70
		SD	2.42	2.11	0.85
		$F_{vs. SR}$	1.01	9.80	
		$p$	.32	.02	
		$\eta^2$		.19	
	30	Mean	5.02	5.19	4.89
		SD	2.69	2.03	0.99
		$F_{vs. SR}$	0.08	0.43	
		$p$	.78	.51	
	31	Mean	5.62	5.17	5.19
		SD	2.75	1.91	1.10
		$F_{vs. SR}$	0.92	0.00	
		$p$	.34	.98	

Note: OD = Online Distributed, OR = Online Regional, and SR = Stand-up Regional. NA = Not Applicable.  $df = 79$  for all one-sample t-tests. All t-test  $p$  values are  $< .0001$ .  $df = (1,234)$  for all F tests.

### Qualifying Set Performance

A repeated measures ANOVA revealed no statistically significant interaction between qualifying set agreement and reading prompt [ $F_{(4,234)} = 0.78, p = .54$ ], so comparisons were made

between training/scoring groups, collapsing across reading prompts. **Table 3** presents the average percent of exact agreement between raters in each group and the consensus scores assigned by rater trainers to the qualifying responses for each training/scoring group. Overall qualifying agreement was less than 2% higher for the stand-up regional group, but neither of the contrasts was statistically significant. With respect to qualifying rate, **Table 3** presents the average qualification rate for the three training/scoring groups, collapsed across the three reading prompts. There are only small differences between the qualification rates of the three groups with the stand-up regional group qualifying at a slightly higher rate under the higher criterion. Inferential statistics were not computed across reading prompts for qualifying rate because of data sparseness (i.e., a singular covariance matrix).

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

Variable	Statistics	OD	OR	SR	
Qualifying Set Agreement	Mean	76%	76%	78%	
	SD	9.19	10.00	8.63	
	$F_{vs. SR}$	1.18	1.56		
	$p$	.28	.21		
Qualifying Rate	1 of 3 sets	Mean	99%	99%	100%
		SD	9.13	9.13	0.00
	2 of 3 sets	Mean	85%	83%	87%
		SD	35.86	38.15	34.14

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

$df = (1,357)$  for all  $F$  tests.

## Reliability & Validity Performance

**Table 4** displays the average of each of the score quality indices along with the standard deviation, and relevant inferential statistics. Preliminary analyses revealed very small differences between the values of the reliability coefficients and the validity coefficients across the items for each group. Specifically, the average reliability coefficients ranged from .70 to .76 and were consistent relative to training/scoring group across items. Similarly, the average validity coefficient ranged from .58 to .64, remaining relatively consistent between the training/scoring groups across the items. Hence, t-tests were conducted on Fisher transformations of the average within-group reliability and validity coefficients. A repeated measures ANOVA indicates that the item-by-group interaction is statistically significant, but the effect size is small [ $F_{(4,234)} = 2.47$ ,  $p = .05$ ,  $\eta^2 = .03$ ]. Therefore, the agreement with consensus scores, reported here as the validity coefficient, is summarized collapsing across items. Among the group comparisons for all three of these rater performance indices, there are no notable between-group differences.

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

Variable	Statistics	OD	OR	SR
Inter-rater Reliability				
	Mean	.75	.73	.72
	$t_{vs. SR}$	0.25	0.09	
	$p$	.40	.47	
Validity Coefficient				
	Mean	.63	.62	.60
	$t_{vs. SR}$	0.21	0.13	
	$p$	.42	.45	
Validity Agreement Index				
	Mean	64%	64%	65%
	SD	3.95	4.58	4.13
	$F_{vs. SR}$	0.01	0.30	
	$p$	.93	.58	

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

$df = 39$  for all t tests.

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

### Backread Agreement

**Table 5** displays descriptive statistics for backreading rate and backreading agreement by training/scoring context group. Backreading rate was somewhat greater in the online regional group than for the other two training/scoring context groups. It is unclear why this trend exists. A repeated measures ANOVA revealed a statistically significant interaction between reading prompt and training/scoring context group with a moderately large effect size [ $F_{(4,230)} = 4.00, p = .004, \eta^2 = .06$ ], so comparisons were made between training/scoring groups for each reading prompts. Backreading agreement rates were highest for the online regional group, and the agreement rates remained fairly consistent across items for both regional groups. Initially, the

agreement rates were lowest for the online distributed group, but they increased to the same level as those observed for the stand-up regional group by the third reading prompt.

**Table 5: Backreading Agreement by Group**

Variable	Item	Statistics	OD	OR	SR	
Backreading Rate	Averaged	Mean	24.30	32.60	22.32	
		SD	3.70	8.39	2.83	
Backreading Agreement	29	Mean	73%	75%	81%	
		SD	11.42	10.28	10.27	
		$F_{vs. SR}$	11.32	8.22		
		$p$	.001	.005		
		$\eta^2$	.07	.05		
	30	Mean	77%	76%	80%	
		SD	11.20	11.06	8.99	
		$F_{vs. SR}$	1.40	2.39		
		$p$	.24	.12		
31	Mean	83%	75%	83%		
	SD	8.54	10.90	11.00		
	$F_{vs. SR}$	0.07	11.36			
	$p$	.80	.001			
	$\eta^2$		.06			

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

$df = (1,115)$  for all F tests.

### Perception of Training and Scoring

**Table 6** displays the average measures from the two rater perception scales for each training/scoring context group. On both scales, the stand-up regional group exhibited the most positive perceptions, and the online regional group exhibited the least positive perceptions. However, none of the group comparisons is statistically significant.

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

Variable	Statistics	OD	OR	SR
Effectiveness	Mean	1.61	1.54	1.73
	SD	0.41	0.43	0.24
	n	49	14	19
	t vs. SR	1.51	1.49	
	df	57	19	
	p	.24	.15	
	Satisfaction	Mean	1.60	1.53
SD		0.43	0.44	0.27
n		49	14	19
t vs. SR		1.55	1.50	
df		53	20	
p		.14	.15	

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

## DISCUSSION & CONCLUSIONS

These analyses indicate that the following differences seem to exist between training/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 and one-half times longer to complete than online training (about 2 hours for online training versus about 5 hours for stand-up training per reading prompt). This is likely because of the human interactions required for stand-up training. Trainers in that context likely spend time introducing themselves, answering questions from individuals, and manipulating materials. These tasks 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 only one statistically significant difference, the amount of time required for stand-up regional scoring was initially slightly less than for the two online contexts. However, these differences were only apparent on the first of the three reading prompts, so they are probably relevant only for projects that are short in duration. It is 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 does not influence the immediate performance of raters following training.*** Our data indicate that agreement rates on qualifying sets were equivalent for

online distributed and online regional raters (76%), and the stand-up regional group performed only slightly better (78%). 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 apparent differences in the performance of raters in the three groups—nearly all the raters in each group achieved that standard. On the other hand, raters in the online regional had a slightly lower qualification rate (83%) than those in the online distributed (85%) and stand-up regional (87%) contexts when adhering to the higher qualifying standard (i.e., 70% or better agreement on at least two of three qualifying sets).

***Third, training/scoring context does not seem to influence rater accuracy.*** None of the observed differences were statistically significant. With respect to inter-rater reliability and the validity coefficient—measures of the consistency of rank ordering of student responses—the online distributed group performed slightly better than the other two groups. On the other hand, in terms of agreement (i.e., the validity agreement index), the three training/scoring context groups performed equally—all achieving about 65% agreement with scoring project leaders. As one might expect, stand-up regional raters were able to attain backreading agreement rates that were initially higher, but these differences diminished after the first reading prompt so that the three groups obtained comparable backreading agreement rates.

***Fourth, scoring context does not seem to influence rater perceptions of the training and scoring process.*** Our analysis of the rater perception measures revealed no statistically significant differences on either rater perception of the effectiveness of or their level of satisfaction with training and scoring materials, procedures, and staff.

It is important to interpret these differences in light of the fact that these data come from intact groups—we were unable to randomize 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 there were no statistically significant differences between groups on any of these variables. Still, 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 more scoring experience (i.e., those in the online regional and stand-up regional groups) 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. On the other hand, one could present arguments favoring the raters in both the distributed and regional contexts based on demographic differences. Research relating to human-computer interactions indicates that younger people (e.g., in the distributed context) tend to show higher levels of facility with computer tasks than do older people (Colley & Comber, 2003). That same body of research has indicated that blacks (e.g., those in the regional context) and females (e.g., those in the distributed context) 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 differences 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 because 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 there are only small differences between the scoring contexts with respect to scoring quality.

## 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.