J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 10, Number 1 · July 2010

The Effectiveness and Efficiency of Distributed Online, Regional Online, and Regional Face-to-Face Training for Writing Assessment Raters

Edward W. Wolfe, Staci Matthews, & Daisy Vickers

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 10, Number 1

The Effectiveness and Efficiency of Distributed Online, Regional Online, and Regional Face-to-Face Training for Writing Assessment Raters

Edward W. Wolfe, Staci Matthews, & Daisy Vickers

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Jennifer Higgins
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free online journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Wolfe, E.W., Matthews, S., & Vickers, D. (2010). The Effectiveness and Efficiency of Distributed Online, Regional Online, and Regional Face-to-Face Training for Writing Assessment Raters. *Journal of Technology, Learning, and Assessment, 10*(1). Retrieved [date] from http://www.jtla.org.

J·T·L·A

Abstract:

This study examined the influence of rater training and scoring context on training time, scoring time, qualifying rate, quality of ratings, and rater perceptions. One hundred twenty raters participated in the study and experienced one of three training contexts: (a) online training in a distributed scoring context, (b) online training in a regional scoring context, and (c) stand-up training in a regional context. 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.

J·T·L·A

The Effectiveness and Efficiency of Distributed Online, Regional Online, and Regional Face-to-Face Training for Writing Assessment Raters

Edward W. Wolfe
Staci Matthews
Daisy Vickers
Pearson

Introduction

Human scoring of responses to constructed-response assessment items has traditionally taken place at regional scoring centers, at which trainers meet and present training materials to raters in a face-to-face setting. Following training, raters receive and review paper copies of the products to be scored and assign scores on handwritten or scannable forms, which are collected and entered into a database. The process of distributing products and collecting and entering scores is somewhat slow, making it difficult to conduct rater monitoring in a timely manner. Over time, technology has been developed to facilitate the processes of training raters, distributing products to be scored, and collecting assigned scores. For example, it is now possible for raters to receive training materials and scanned copies of products to be scored through a computer interface and for scoring project directors to collect scores that raters enter into a graphical computer interface. As a result, raters can conceivably be trained, qualify for scoring, and assign scores from remote locations, such as their homes, without ever meeting with scoring project directors in a face-to-face setting.

Few research studies have focused on how features of the training and scoring context, such as computer-based distribution and collection systems, affect the quality of scores assigned by human raters. Given that the reliability of scores of constructed-response items tends to be low, relative to scores from multiple-choice items, test developers continuously seek ways to eliminate error from scores assigned to constructed-response items. Typically, these efforts include rater training, requiring raters to qualify for a scoring project, monitoring and retraining raters during the

scoring project, and utilizing score resolution procedures when the scores of two or more raters differ. This article 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* training 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* training 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).

Much of the initial research concerning rater training focused on training content: (a) rater error training, in which raters are informed of the existence of rating errors (e.g., leniency, assigning scores that are lower than valid scores, and halo, assigning scores that are too highly related across multiple items) and how to avoid those errors, and (b) frame of reference training in which raters learn about relevant aspects of the products to be rated. That research revealed that rater error training may reduce the occurrence of leniency and halo errors better than frame of reference training (Ivancevich, 1979), whereas assigned ratings may be more accurate when frame of reference training or no training at all is provided to raters (Noonan & Sulsky, 2001; Roch & O'Sullivan, 2003; Uggerslev & Sulsky, 2008). Several of these initial studies revealed that either approach to rater training is more effective when training immediately precedes rating (Roch & O'Sullivan, 2003; Sulsky & Day, 1994). One study suggested that each approach may be better at controlling specific types of errors (Stamoulis & Hauenstein, 1993). However, a combination of rater error training and frame of reference training was shown to be superior to either rater error training or frame of reference training alone (McIntyre, Smith, & Hassett, 1984; Pulakos, 1984). Finally, at least one study revealed that rater reactions to training may influence the effectiveness of training efforts (Noonan & Sulsky, 2001).

Few systematic studies have compared the effectiveness or efficiency of online and face-to-face rater training contexts. In one of the few direct comparisons of online and face-to-face rater training, Knoch, Read, and von Randow (2007) compared the performance and attitudes of two teams of eight writing assessment raters. Their results revealed that both training formats reduce rater severity, inaccuracy, and central tendency. Moreover, raters in the online training condition were more similar to one another in terms of levels of these rater effects following training than were those in the face-to-face condition. The authors found no differences in terms

of rater perceptions or preferences for the two training media. A pair of studies (Elder, Barkhuizen, Knoch, & von Randow, 2007; Elder, Knoch, Barkhuizen, & von Randow, 2005) followed eight experienced raters who rated writing samples online before and after receiving online training concerning the rating task. During training, raters rated writing samples and received immediate feedback through that interface concerning the accuracy of the scores that they assigned. The raters generally exhibited positive attitudes toward the online training system, indicating that the system was effective and enjoyable and that the online training system changed the raters' behaviors. The analyses also revealed that the range of rater severity decreased after completing the online training.

The purpose of this study is to directly compare the effectiveness and efficiency of online distributed, online regional, and face-to-face (stand-up) regional training of writing assessment raters. Specifically, we seek to determine whether the speed of training, the quality of assigned scores, and the perceptions of raters engaged in these three training and scoring contexts are comparable. Our research addressed the following questions.

1. Do raters who undergo online distributed, online regional, and stand-up regional training produce scores that differ in terms of psychometric quality?
2. Do raters who undergo online distributed, online regional, and stand-up regional training complete the training and scoring processes at different rates of speed?
3. Do perceived effectiveness of and satisfaction with training experiences vary across raters completing online distributed, online regional, and stand-up regional training?

Method

To address these research questions, we conducted a quasi-experimental design which employed explicit matching of 120 raters who were distributed between three training conditions (40 raters in each condition)—online distributed, online regional, and stand-up regional training. Each rater scored a set of 400 secondary student essays that were composed in response to a state-wide writing assessment, using an online distribution system on a four-point rating scale after receiving rubric-specific training. Data were collected concerning the quality of the scores assigned by the raters as well as the amount of time required to complete training and scoring. 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.

Raters

Raters for each group were selected through purposive sampling from a pool of experienced writing assessment raters so that the three groups were comparable in terms of 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 using the rubric on which they were trained, and all participants were paid an equal lump sum for completing the training and scoring.

The demographic, educational, and professional frequencies indicate that the three training context groups were comparable. With respect to demographic characteristics, participants in the online distributed group were slightly more likely to be female (58% versus 47% in the other two groups) and under the age of 55 (68% versus 40% in the other two groups), whereas participants in the online regional and stand-up regional groups were more likely to be white (88% versus 63% in the online distributed group). 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 $\chi^2_{(8)\text{Ethnicity}}=13.73$, $p=.05$. Concerning educational experiences, the online distributed raters were more likely to have attained a graduate degree than the other two groups (45% versus 20% for the other two groups), although the difference was not statistically significant [$\chi^2_{(4)}=7.61$, $p=.13$]. On the other hand, we did observe a statistically significant difference concerning undergraduate major, and these frequencies are displayed in Table 1 (next page). Specifically, the online distributed raters were more likely to provide no response to undergraduate major than the other two groups, [$\chi^2_{(8)}=22.07$, $p=.006$]. Finally, with respect to teaching experience, the online distributed raters were more likely to have previously participated in four or more scoring projects than were raters in the other two groups (73% versus 63% in the other two groups) and the online regional raters were less likely to have secured a teaching certification (83% versus 73%). 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: Demographics, Education, and Experience by Training Context

Variable	Level	Online Distributed	Online Regional	Stand-up Regional
Undergraduate Major	Business	23% (9)	35% (14)	38% (15)
	Humanities/Lib Arts	50% (20)	53% (21)	46% (23)
	Sciences	8% (3)	33% (5)	5% (2)
	No Response	20% (8)	0% (0)	0% (0)

Note: Percentages (and frequencies) for each group are displayed for each level of each variable.

Materials and Procedures

Raters were trained to apply a four-point, focused, holistic scoring rubric using training materials that were originally developed for stand-up training that was delivered to a different group of raters who participated in an operational scoring project from which the student essays in this study were sampled. Members of the range-finding committee from the operational project assigned consensus scores to responses which were compiled into two sets of ten practice papers (completed by raters during training) and three sets of ten qualifying papers (scored by raters at the conclusion of training but prior to scoring). At least three members of the range-finding committee members independently scored each of 600 randomly-selected student responses, and then the committee members jointly reviewed the responses to arrive at consensus scores. They also reviewed the papers to determine whether any were off-topic, unusual, illegible, or could not receive a numeric score. The scoring directors then worked together to choose the 400 responses raters in the study would score; they were instructed 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. In operational scoring, when scorers encounter responses that meet these criteria, they are instructed to send the response to a scoring supervisor, rather than attempting to score the responses. Because of this, scoring directors excluded such responses from the sample. Scoring directors also excluded any responses that were so unusual that they were not at all representative of the responses seen in live scoring. Responses were not excluded simply because they were more difficult to score or not "clear cut" responses. From those remaining, 400 responses were selected for use in the study. Scoring directors also selected 16 additional calibration (ongoing training) essays, which were seeded randomly into all raters' scoring queues and for which raters received feedback concerning the accuracy of the scores that they assigned to these essays.

A content specialist, familiar with online and stand-up training, reviewed the training materials and made adjustments for online training. The content specialist arranged the 30 training papers into three sets of ten papers each. That configuration was used for all groups in the study. The content specialist also reviewed and edited annotations for grammar and content consistency. During the stand-up training portion of the study, the scoring directors used these annotations as the basis for content explanations of anchor and training papers. The scoring directors completed these online training modules and online practice and qualification sets prior to the study and provided feedback to increase the comparability of the online and stand-up training materials and procedures. With the exception of the fact that those participating in online training viewed images of the original response, whereas 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 wording (beyond what was presented in the online training) in the stand-up training group. These standardized annotations were presented during online training for each example essay. 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 informally documented questions and interventions asked by raters during training and scoring.

At the end of training, all raters rated three sets of ten essays for which had been assigned consensus scores—a process referred to as "qualifying" in operational projects. In all three groups of raters, 98% of the raters attained the rate of agreement (70% perfect agreement) on at least one of the three qualifying sets—a standard that is typical of operational projects (for example, see page 25 of http://ritter.tea.state.tx.us/student.assessment/ELL/telpas_rater_user_guide.pdf). Although raters who fail to attain this standard in operational settings do not assign ratings reported to students, those raters were permitted to assign ratings in this research project. The noteworthy point, for this study, is that the number of raters who would not have qualified after training was small and was consistent across the three training context groups.

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 performance on the scoring task, and rater perceptions of the effectiveness of and their satisfaction with the training and scoring context they experienced.

Time

Scoring and training times 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

Reliability and validity were measured in three 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, an index that is not sensitive to rater severity or leniency. 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, another index that is not sensitive to rater severity or leniency. Third, the *validity agreement index* was defined as the percentage of exact agreement between the scores assigned by raters to the 400 essays and the consensus scores assigned by project leaders—an index that is influenced by several rater effects (e.g., severity/leniency, centrality/extremism, and accuracy/inaccuracy).

Rater Perceptions

Rater perception of training and scoring procedure effectiveness, as well as 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 and materials, personnel, qualifying process, scoring process, scoring materials, etc.). Coefficient alpha for the effectiveness and satisfaction scales equals α=.95 and α=.96, respectively. The correlation between effectiveness and satisfaction scale scores equals .59. A copy of the questionnaire is shown in the Appendix.

Analyses

In our analyses, training context was treated as an independent variable, and the analyses focused on determining whether training groups differed on each outcome variable. In most cases, planned comparisons were conducted in which the *online distributed* and the *online regional* raters were compared to 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 δ for each *t*-test and η^2 for each Analysis of Variance (ANOVA).

Time

A one-sample *t*-test was conducted to determine whether the *online distributed* and the *online regional* number of training hours differed from the constant number of hours the *stand-up regional* group spent in training. An ANOVA was conducted to determine whether the training context groups differed with respect to the number of hours spent scoring.

Rater Performance

T-tests were conducted to evaluate training context group differences on Fisher transformations of the inter-rater reliability and validity coefficients, and an ANOVA was conducted to determine whether the training context groups differed on an arcsine transformation of validity agreement index (measured as a percentage).

Rater Perceptions

An ANOVA was conducted to determine whether the training 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 the raters participating in this (writing) study and a different group of raters participating in a companion study focusing on reading (Wolfe, Matthews, & Vickers, 2009).

Results

Overall, the results indicated no differences between the training context groups with respect to score quality, scoring time, or perceptions of training and scoring. However, training took considerably less time for raters in the two online training contexts when compared to those in the stand-up training context.

Reliability and Validity Performance

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 context group are shown in Table 2. Overall, the online distributed group assigned ratings of slightly higher quality in comparison to the ratings assigned by the two regional training groups. However, none of the observed differences are 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 2: Reliability and Validity Performance by Training 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
	z 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 validity agreement. *df*=(1,119) for the *F* tests.

Training and Scoring Time

The number of training and scoring hours for each training context group are summarized in Table 3. The number of hours of training for the stand-up regional group is 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 is statistically significant, although the online distributed raters took slightly less time to finish scoring the 400 essays.

Table 3: Scoring and Training Hours by Training Group

Variable	Statistics	OD	OR	SR
Training Time	Mean	3.40	4.75	12.00
	SD	1.40	1.37	
	t vs. SR	38.84	33.52	
	δ	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. For the t-tests, df=39 and p values are < .0001. For the F tests, df=(1,119). All SR raters completed training in the same session, so there is no within group variability for this condition.

Perception of Training and Scoring

The average score on the two rater perception scales for the three training context group is shown in Table 4. On both scales, measures for the online distributed and stand-up regional groups are slightly higher than those of the online regional group. However, none of these differences is statistically significant.

Table 4: Training and Scoring Perception Measures by Training 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).

Anecdotally, a review of supervisor logs suggests 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, whereas online regional raters seemed more likely to request assistance for issues relating to the online training and scoring computer interface.

Discussion

These results suggest several points concerning the nature of online and stand-up training and the distributed and regional scoring contexts. *First, training takes less time when delivered online, but online scoring time is not greatly affected by context.* With respect to training time, stand-up training as delivered in this study took about three times longer to complete than online training (about four 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 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 underestimate 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" are 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 may significantly underestimate the number of calendar days that would be required for those raters to complete the scoring project (e.g., those raters could log only a minimal amount of time per day, resulting in a higher number of calendar days to complete the project). However, this is speculation on our part. One can easily argue that distributed scorers could spend more time on task due to time saved commuting to a regional center and being able to work at odd hours that would not be afforded to those in regional centers.

Second, the training context does not seem to influence the quality of ratings. In this study, raters in the online distributed context were slightly better able to agree with one another (i.e., higher inter-rater reliability) and were slightly better able to agree with scoring leaders (i.e., higher validity coefficients). However, none of the observed differences are statistically significant. Hence, it seems that the online training context is more efficient than the stand-up context—it took less time and produced equivalent rater accuracy. Not only did raters in the online groups complete training in about one-third the time as stand-up groups, but the scores that they assigned were of equal quality to those assigned by raters who experienced

stand-up training. This outcome is consistent with the results of a study by Knoch, Read, and von Randow (2007) who found that raters who experienced online training were less likely to exhibit rater effects.

Third, training context does not influence rater perceptions of the training and scoring process. In this study, raters who experienced the online regional training context had slightly 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 groups. However, the observed differences were not statistically significant. This outcome is consistent with the results of previous studies that found no difference between the perceptions of raters trained in online and face-to-face contexts (Knoch, et al., 2007) and generally positive attitudes of raters trained in online contexts (Elder, et al., 2007; Elder, et al., 2005).

Conclusions

It is important to interpret these differences in light of the fact that these data come from groups that were constructed to be comparable. In a real-world setting, the populations from which distributed and regional raters are recruited may be very different. For example, regional rater populations will have demographic, educational, and experiential characteristics that reflect those of qualified individuals who are in proximity to the scoring center or those who can travel to the scoring center. Those who can participate in a distributed context, on the other hand, will likely be drawn from a larger and more widely distributed population. It is also possible that the requirement that distributed training and scoring takes place on a computer could truncate the potential population of those available to score in that context. Because of these population-level differences, it was necessary to purposively sample groups in order to make them as comparable as possible on relevant background variables.

This fact limits the generalizability of our results in two ways. First, it is possible that the small differences between the groups with respect to demographic, educational, and experiential variables may have influenced the outcome variables that we chose to study. We believe that this would only be an important concern if we had observed important group differences on those variables. However, the only outcome that produced a statistically significant difference between the training context groups was training time. We believe that this fact, coupled with the intentional similarities that we created between the three groups with respect to background variables, makes a very compelling argument in favor of the efficiency of online distributed rater training and scoring. Second, and somewhat more restricting, is the fact that the selected samples probably are not representative of the populations from which raters would be

drawn in operational scoring projects. Because we sought to create comparability between our comparison groups, we are unable to determine whether (a) the populations from which raters would be drawn for distributed and regional scoring projects are similar to one another and (b) the online and stand-up training procedures would produce comparable results with those two distinct populations.

What we did learn, however, is a valuable piece of information. Given a common population of raters, online rater training, as implemented in the system employed in this study, is more efficient (i.e., takes less time to produce comparable results) than stand-up training. This outcome is important because it offers several potential advantages to those conducting scoring projects. First, online distributed training avoids the logistic difficulties associated with convening a panel of raters in one location. Automated training can be scheduled at times convenient for the rater, and raters do not have to travel to a common location to receive training.

Second, online distributed training offers improved opportunities for individualizing training for a particular rater's needs. Although we did not utilize these technologies in our study, it is possible to provide remedial training to raters who display evidence of specific idiosyncrasies or errors in the scores that they assign. Similarly, it is possible to take advantage of user interface features that could potentially improve the effectiveness of online distributed training—again, a feature that was not implemented in this study.

Third, online distributed training and scoring may make a broader population available for participation in a particular operational scoring project. The raters could be chosen without regard to geographic location or mobility, which is a restriction faced in a regional scoring context. In addition, online distributed options may open the population of raters up to those who have irregular schedules that would make it difficult to work in scoring projects held during fixed working hours. The fact that online rater training appears to be more efficient than stand-up training makes it clear that those conducting rater training can safely pursue these opportunities provided by an online distributed scoring context.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37–64.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*(3), 175–196.

Ivancevich, J.M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. *Journal of Applied Psychology, 64*(5), 502–508.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26–43.

McIntyre, R.M., Smith, D.E., & Hassett, C.E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology, 69*(1), 147–156.

Noonan, L.E., & Sulsky, L.M. (2001). Impact of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance, 14*(1), 3–26.

Pulakos, E.D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69*(4), 581–588.

Roch, S.G., & O'Sullivan, B.J. (2003). Frame of reference rater training issues: recall, time and behavior observation training. *International Journal of Training & Development,* 7(2), 93–107.

Stamoulis, D.T., & Hauenstein, N.M. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology, 78*(6), 994–1003.

Sulsky, L.M., & Day, D.V. (1994). Effects of frame-of-reference training on rater accuracy under alternative time delays. *Journal of Applied Psychology, 79*(4), 535–543.

Uggerslev, K.L., & Sulsky, L.M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology, 93*(3), 711–719.

Wolfe, E.W., Matthews, S., & Vickers, D. (2009). *A comparison of training & scoring in distributed & regional contexts-reading* (Pearson Research Report). Iowa City, IA: Pearson.

Appendix

Rater Questionnaire

Instructions:
Please indicate your evaluation of how *effective* each of the following aspects of the scoring project were as well as your *satisfaction* with that aspect of the scoring experiences.

	Effectiveness			Satisfaction		
Aspect	Not very effective	Moderately effective	Very Effective	Not very effective	Moderately effective	Very Effective
Training procedures	o	o	o	o	o	o
Training materials	o	o	o	o	o	o
Personnel who conducted training	o	o	o	o	o	o
Qualifying procedures	o	o	o	o	o	o
Qualifying materials	o	o	o	o	o	o
Personnel who conducted qualifying process	o	o	o	o	o	o
Scoring process	o	o	o	o	o	o
Scoring materials	o	o	o	o	o	o
Personnel who conducted scoring process	o	o	o	o	o	o
Feedback provided to you concerning your performance as a rater	o	o	o	o	o	o
Recalibration materials	o	o	o	o	o	o
Recalibration process	o	o	o	o	o	o
Personnel who provided feedback to you and conducted recalibration	o	o	o	o	o	o
Availability of information to clarify questions that arose during scoring	o	o	o	o	o	o
Quality of information to clarify questions that arose during scoring	o	o	o	o	o	o
Personnel who answered questions that arose during scoring	o	o	o	o	o	o

Author Biographies

Edward W. Wolfe is a Senior Research Scientist in the Assessment & Information group of Pearson. In that position, Dr. Wolfe conducts research in support of test development and scoring. Dr. Wolfe's research focuses on applications of latent trait models to instrument development, the analysis of ratings, and computer-based testing. Dr. Wolfe received his Ph.D. in Educational Psychology with an emphasis in the areas of Psychometrics and Cognitive Science from the University of California at Berkeley in 1995. He received his M.S. in Educational Psychology from Purdue University in 1990, and he received his B.A. in Music/Education from Fairmont State College in 1987. Edward Wolfe can be reached at ed.wolfe@pearson.com.

Staci Matthews is the Manager of Content, Quality, and Training for Performance Scoring in the Assessment & Information group of Pearson. In that position, Ms. Matthews implements training programs for content staff and works with scoring content staff and customers on scoring design, training, and quality issues as well as supporting research for performance scoring. Prior to her current position, she worked as a program manager for scoring programs. She received her B.A. Degree in English from the University of Iowa. Staci Matthews can be reached at staci.matthews@pearson.com.

Daisy Vickers is a Director of Performance Scoring in the Assessment & Information group of Pearson. In that position, Ms. Vickers consults with clients on scoring models and quality issues. Within Pearson, she consults with Program Management and Content Management on scoring design and quality across all projects as well as supporting research for performance scoring. Before joining Pearson, she was Chief Consultant of Performance Assessment with the North Carolina Department of Public Instruction for thirteen years. She received her M.A. Degree in English and her B.A. in English/Education from East Carolina University. Daisy Vickers can be reached at daisy.vickers@pearson.com.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College