Evaluation of Pseudo-Scoring as an Extension of Rater Training

Research Report

Edward W. Wolfe
Melodie Jurgens
Bob Sanders
Daisy Vickers
Jessica Yue

April 2014

ALWAYS LEARNING PEARSON

About Pearson
Everything we do at Pearson grows out of a clear mission: to help people make progress in their lives through personalized and connected learning solutions that are accessible, affordable, and that achieve results, focusing on college-and-career readiness, digital learning, educator effectiveness, and research for innovation and efficacy. Through our experience and expertise, investment in pioneering technologies, and promotion of collaboration throughout the education landscape, we continue to set the standard for leadership in education. For more information about Pearson, visit http://www.pearson.com/.

About Pearson's Research Reports
Our network **mission** is to spark innovation and create, connect, and communicate research and development that drives more effective learning. Our **vision** is students and educators learning in new ways so they can progress faster in a digital world. Pearson's research papers share our experts' perspectives with educators, researchers, policy makers and other stakeholders. Pearson's research publications may be obtained at: http://researchnetwork.pearson.com/.

Abstract

This report summarizes the results of a study that sought to determine the potential benefit of engaging raters of essays written by students in a pseudo-scoring process as an extension of rater training (i.e., practice scoring that occurs between training and operational scoring). Raters at three grade levels took part in rater training activities, then took a pre-qualifying test, then engaged in pseudo-scoring, then took a qualifying test. Those who achieved qualifying status then proceeded to operational scoring. Results indicate an increase in performance on qualifying sets, an increase in qualifying rate, and an increase in inter-rater correlation following pseudo-scoring. Implications for operational scoring projects are provided.

Keywords: Performance Assessment, Raters, Scoring, Training

Evaluation of Pseudo-Scoring as an Extension of Rater Training

Scoring projects that involve human raters, particularly in large-scale performance assessments, seek different methods to minimize the amount of measurement error associated with assigned scores. First, the rating process is carefully designed to standardize the scoring criteria and process so that all raters are introduced to the criteria and conduct the rating process in a similar manner. Second, typically raters are engaged in extensive training prior to operational scoring, using standardized training materials and procedures, in order to communicate the intent of the leaders of the scoring project concerning appropriate interpretation and application of the scoring rubric and process. Third, raters are typically required to exhibit and maintain a predetermined level of mastery of the scoring rubric by meeting qualification criteria prior to operational scoring and by exhibiting adequate performance during subsequent monitoring of their operational performance. Although some prior research exists concerning the efficacy of rater training, one topic that has not been addressed in prior studies is the degree to which practice applying the scoring rubric, referred to in this report as pseudo-scoring, facilitates rater mastery of a scoring rubric. This report summarizes an empirical study that seeks to address this issue.

Woehr and Huffcutt (1994) reviewed much of the literature regarding rater training in the context of performance appraisal, and that review and subsequent meta-analysis summarizes the impact of rater training in terms of four training foci: (a) rater errors (i.e., training raters to recognize and avoid rating errors, such as leniency, centrality, and halo), (b) performance dimensions (i.e., training raters to recognize and base decisions on appropriate characteristics of the performance), (c) frame-of-reference (i.e., training raters with respect to performance standards as well as performance dimensions), and (d) behavioral observations (training raters to effectively observe performances). Of these four approaches, frame-of-reference training is the focus most commonly adopted in large-scale operational performance assessment scoring in the US. Woehr and Huffcutt's meta-analysis revealed several trends. First, rater error training appears to be effective in reducing several types of rater effects, particularly when that training focuses on helping raters understand and avoid common rater errors. Second, performance dimension training seems to be most effective in minimizing halo effects. Third, frame-of-reference training appears to be the most effective training strategy with respect to improving rating accuracy. Fourth, behavioral observation training results in a moderate improvement in rating accuracy. The results concerning frame-of-reference training have been replicated in the context of essay scoring by Weigle et.al. (1994, 1998, 1999), who also demonstrated that this training method causes inexperienced raters to perform more similarly to experienced raters following training. Shohamy, Gordon, and Kraemer (1992) also replicated these results and found that rater training may also be effective regardless of raters' professional backgrounds (e.g., professional training versus lay).

More recently, in an effort to improve the rater training, some researchers have studied the impact of online training. This line of research has suggested that although the medium through which rater training is delivered may have only a small impact on rater performance, online training is more efficient than traditional face-to-face rater training (Elder, Barkhuizen, Knoch, & von Randow, 2007; Knoch, Read, & von Randow, 2007; Wolfe, Matthews, & Vickers, 2010). It is also worth noting that providing more individualized rater feedback via an online rating interface may reduce the occurrence of rater disagreement (Elder et al., 2007). In addition, raters have been shown to differ in their receptivity to online rater training (Elder et al., 2007; Hamilton, Reddel, & Spratt, 2001; Knoch et al., 2007).

A different practice that may improve the performance of raters in operational projects, the one that is the focus of the current study, is providing them with practice applying the scoring rubric in a realistic context. We refer to this practice as pseudo-scoring. Unfortunately, little, if any, research has been directed toward determining the efficacy of pseudo-scoring practices on rater performance. However, based on the considerable research regarding the role of practice in the development of expertise, we would expect practice of complex cognitive tasks, such as engaging raters in pseudo-scoring, should improve their performance. Specifically, Ericsson and his colleagues (Ericsson & Charness, 1994; Ericsson, Krampe, & Tesch-Romer, 1993; Ericsson, Prietula, & Cokely, 2007) conclude that it is not innate ability that separates experts from everyone else. Rather, it is extended and deliberate, thoughtful practice of unmastered skills over an extended period of time, perhaps years, that helps one develop expertise in a particular area of performance. This report attempts to examine this issue as it relates to professional scoring by summarizing the results of a study that monitored the performance of raters before and after they engaged in pseudo-scoring in an operational scoring project for the purpose of determining the potential contribution that pseudo-scoring makes to helping raters master a scoring rubric beyond the level attained during a standard rater training process. Specifically, we address the following research questions.

- Are raters more likely to achieve qualifying status following pseudo-scoring?
- Does the quality of assigned scores improve following pseudo-scoring?
- Is the impact of pseudo-scoring consistent across rater characteristics groups?

Method

We collected data from 713 raters who participated in an operational scoring project in which a six-point holistic scoring rubric was applied to essays written by approximately 29,000 students across three grade levels (4, 8, and 10). Raters were assigned to grade levels based on the raters' physical location (one of three cities across the United States) and then underwent a three-day training process for their grade-specific rubric using grade-specific training materials. At the conclusion of training, each rater undertook a pre-qualifying test by assigning scores (blindly) to essays that had been previously assigned consensus scores by training leaders. Following the pre-qualifying test, each rater engaged in two days of pseudo-scoring and received feedback on their scoring performance in the form of periodic statistical reports, generated by a computer-based scoring system and verbal communications, provided by scoring trainers. Trainers used statistical summaries and periodic reviews of scores assigned (i.e., backreading) as the basis for their feedback. At the end of the two days of pseudo-scoring, raters took a formal qualifying test. The test comprised three sets of 20 essays each. Raters who achieved an average of 70% perfect agreement and 100% perfect-plus-adjacent agreement on the first two qualifying or who achieved 70% perfect agreement on the best two of three qualifying sets with no more than one non-adjacent rating across all three qualifying sets were permitted to begin operational scoring. Raters who attained neither of these standards were released from the scoring project.

Raters

Table 1 summarizes the demographic characteristics of the raters at each grade level. Over 200 raters participated at each of the three grade levels, and, on average, over half of the raters were female and Caucasian with an average age of about 50 years. There were few missing observations for gender and age, but no ethnicity data were available for approximately 20% of the raters due to nonresponses to the demographics questionnaire.

Table 1. Rater Demographics

Gender	**Grade**	**Female**	**Male**	**Missing**
	4	50%	44%	5%
	8	56%	40%	5%
	10	50%	45%	5%
Ethnicity	**Grade**	**Caucasian**	**Non-Caucasian**	**Missing**
	4	59%	18%	23%
	8	62%	20%	18%
	10	65%	15%	20%
Age	**Grade**	**All**	**Missing**	
	4	M = 50.13	5%	
		IQR = 39 / 62		
	8	M = 44.19	5%	
		IQR = 29 / 60		
	10	M = 54.82	5%	
		IQR = 49 / 64		

Note: $N_{\text{Grade 4}}$ = 211, $N_{\text{Grade 8}}$ = 232, $N_{\text{Grade 10}}$ = 270. M = Mean. IQR = Interquartile Range. Percentages are conditional on grade and may not sum to 100% within a row due to rounding error.

Table 2 summarizes the professional experiences of the raters at each grade level. A bachelor's degree was the highest level of education for a majority of the raters. At all grade levels, the most commonly reported college majors were in the arts and humanities. The remaining majors reported included business, education, the natural and the social sciences—at least 10% of the raters reported one of these college majors at each grade level. Just over half of the raters reported teaching experience, and over half of the raters also reported prior professional scoring experience. There were very few missing observations with respect to raters' professional experiences.

Table 2. Rater Professional Experience

Education	**Grade**	**Bachelor**	**Master**	**Missing**
	4	72%	28%	0%
	8	74%	26%	0%
	10	70%	30%	0%
Scoring	**Grade**	**Yes**	**No**	**Missing**
	4	36%	64%	0%
	8	49%	50%	0%
	10	46%	54%	0%
Teaching	**Grade**	**Yes**	**No**	**Missing**
	4	48%	52%	0%
	8	42%	58%	0%
	10	44%	57%	0%

Note: $N_{\text{Grade 4}}$ = 211, $N_{\text{Grade 8}}$ = 232, $N_{\text{Grade 10}}$ = 270. Percentages are conditional on grade and may not sum to 100% within a row due to rounding error.

Rubric/Essays

The six-point holistic writing rubrics used by raters to apply scores were unique by grade and were written in prose (not bulleted) format so as to discourage use of the rubric as merely a checklist. Those rubrics generally focus on the essay's purpose, organization and use of transitions, development and support of ideas, and the command of language, including word choice, sentence structure and fluency, and control of grammar and conventions. Students within a grade level responded, in handwriting, to a single prompt that asked students first to consider a topic and then write for a specific purpose. In this study, 4th grade students responded to a narrative prompt, 8th grade students responded to a persuasive prompt, and 10th grade students responded to an expository prompt. Students were given two pages of lined paper on which to write, and they were allocated 60 minutes in which to respond to the prompt. Student responses ranged from a few sentences in length to the full two pages. Responses then were collected, scanned, and loaded into the image-based scoring system.

Training/Qualifying/Scoring

All training and qualifying materials were compiled by experienced content staff via a rigorous process directed and supervised by both a content specialist and state department of education personnel, who ultimately reviewed and provided feedback and approval of all training materials (responses, annotations, and any other project-specific information provided to raters). Raters received synchronous, instructor-led training in a group setting, at regional scoring centers using hard-copy materials for all training activities. Upon completion of training, raters responded to two sets of twenty student responses that were compiled into a pre-qualifying test. These responses were chosen to be representative of the range of student responses and to represent the range of variability in terms of being clear versus borderline examples of each score category. Raters were aware that their responses to the pre-qualifying test would not determine their eventual qualifying status, but they were encouraged to do their best on the pre-qualifying test.

Upon completion of the pre-qualifying test, raters received asynchronous modular training on the scoring system before moving into pseudo-scoring within the same system. All student responses were randomly and anonymously distributed to raters such that they had no indication of student demographics. Responses that were scored during pseudo-scoring were operational student responses that were routed into a scoring queue for the sole purpose of this practice scoring activity; scores applied in pseudo-scoring were not utilized for final score calculation, and raters were aware of this fact. Raters were engaged in the pseudo-scoring activity for approximately 12 hours, although all of that time was not spent scoring student responses. On average, raters scored approximately 91 essays during this period (although the actual number varied considerably, with a standard deviation of 37 essays, ranging from a minimum of 39 essays to a maximum of 279 essays), in addition to scoring validity papers and engaging in calibration training (see below). At the conclusion of pseudo-scoring, raters responded to a qualifying test of the same length and of comparable content to the pre-qualifying test. Raters were aware of the fact that they would not be permitted to continue to operational scoring unless they achieved at least 60% perfect agreement and no less than 90% perfect plus adjacent agreement with consensus scores.

During pseudo-scoring and operational scoring, raters were periodically engaged in calibration training (i.e., a follow up to the initial training process that involves reading and scoring student responses that have been assigned consensus scores and have been annotated by scoring trainers). In addition, raters received feedback concerning their scoring performance in the form of statistical indicators of their agreement levels with other raters as well as with validity responses (i.e., annotated student responses scored by scoring trainers prior to pseudo-scoring). In addition, based on inter-rater reliability (i.e., agreement with scores assigned to student responses by other raters), scoring trainers provided direct feedback to raters based on reviews of scores assigned by raters to individual student responses. All of these activities were carried out during pseudo-scoring and subsequently during operational scoring.

Analysis

Analyses were conducted to identify changes in qualifying and scoring performance that occurred between the end of rater training and the end of pseudo-scoring and to determine whether those changes varied across rater characteristics. That is, we employed a within subjects repeated measures design to determine the impact of pseudo-scoring on rater performance. To this end, we employed statistical comparisons on each of several rater performance outcome variables, using rater characteristics as potential moderators of the observed impact of pseudo-scoring on those outcome variables.

Variables. All analyses were conducted separately by grade level due to the fact that essay prompts, scoring rubrics, and training materials were grade-specific and that grade levels were assigned to unique scoring locations. Hence, the three grade levels can be thought of as three replications of the study's design. We studied pseudo-scoring as the independent variable in our study and considered several potential moderators on the impact of pseudo-scoring on each of several outcome variables. Concerning moderator variables, we considered (a) gender, (b) age, (c) ethnicity, (d) scoring experience, (e) teaching experience, and (f) highest degree. We evaluated the impact of these moderators on the efficacy of pseudo-scoring with respect to the outcome variables that are listed and defined in Table 3, with the exception of IRCorrChange. Specifically, we examined the change in the percent of perfect agreement between the pre-qualifying and qualifying tests (%QGain), and comparability of qualification status between the pre-qualifying and qualifying tests (QStatusChange) based on the criteria that raters must achieve a 65% or better perfect agreement and 90% or better adjacent agreement. We also examined the change in the value of the inter-rater correlation for raters who achieved qualifying status between their pseudo-scores and operational scores (IRCorrChange).

Table 3. Outcome Variables

Variable	Definition	Description
%QGain	% Qualifying perfect agreement gain	Increase/decrease in % of perfect agreement between pre-qualifying and qualifying tests
QStatusChange	Qualifying status improvement	Change in attainment of qualifying status on pre-qualifying and qualifying tests (fail if % perfect < 65% or % adjacent < 90%)
%IRCorrChange	Inter-rater correlation improvement	Percentage of qualified raters for whom the operational score inter-rater correlation is greater than that correlation for pseudo-scores

Analysis Procedures. For analyses involving %QGain, we conducted a simple t-test on the percentage increase at each grade level, and we conducted an Analysis of Variance (ANOVA) to determine whether these increase scores varied across each rater moderator variable separately for each grade level. For analyses involving QStatusChange, we conducted a χ^2 test separately at each grade level to determine whether the pre-qualifying qualification status percentage was equivalent to the final qualifying percentage. Further, we conducted another set of χ^2 tests separately at each grade level to determine whether improvement between pre-qualifying status and final qualifying status (i.e., transitioning from a "fail" on the pre-qualifying test to a "pass" on final qualifying test) varied across levels of each rater background variable. Finally, for analyses involving IRCorrChange, we conducted a separate statistical test for each rater to determine the equivalence of the inter-rater correlations from pseudo-scoring and the inter-rater correlation from scoring for those raters who achieved qualifying status on the final qualifying test. Specifically, we conducted a Fisher transformation of those correlations and then conducted a Z test on the transformed correlations. We report the percentage of raters who exhibited a statistically significant increase in inter-rater correlations at each grade level.

Results

Qualifying

Table 4 summarizes the increase in the percent of perfect agreement between the pre-qualifying and qualifying tests (%QGain) by grade level. Initial performance on the pre-qualifying test was comparable at Grades 4 and 8, but performance on that test was slightly lower at Grade 10. Raters at Grade 4 exhibited only a small increase in the percent following pseudo-scoring (less than 1%). However, larger and statistically significant gains were realized by Grade 8 and Grade 10 raters (10% and 4%, respectively).

Table 4. Qualifying % Perfect Agreement by Grade Level

Grade	Statistic	Pre-Qualifying %	Qualifying %	%QGain
4	**Mean**	60.84	61.47	0.14
	IQR	50.0 / 70.0	55.0 / 72.5	−10.0 / 12.5
8	**Mean**	60.37	70.64	10.05*
	IQR	50.0 / 70.0	62.5 / 80.0	−7.5 / 10.0
10	**Mean**	53.66	57.74	3.96**
	IQR	45.0 / 65.0	47.5 / 70.0	−7.5 / 15.0

Note: $N_{\text{Grade 4}} = 211$, $N_{\text{Grade 8}} = 232$, $N_{\text{Grade 12}} = 270$. IQR = Interquartile Range. * $t_{(231)} = 10.83$, $p < .0001$. ** $t_{(269)} = 3.96$, $p < .0001$.

Subsequent analyses of qualifying gains focused on the stability of these gains across rater demographic and professional experience groups. No statistically significant differences were observed at the three grade levels for Gender, Ethnicity, Teaching Experience, or Highest Degree. Analysis of rater Scoring Experience produced a statistically significant difference at Grades 8 [$F_{(1,230)} = 7.80$, $p = .006$, $R^2 = .03$], but not at Grades 4 and 10. Gains for experienced scorers at Grade 8 were about 5% greater than gains for non-experienced scorers (13% versus

8%, respectively). Analysis of rater Age produced a statistically significant relationship at Grade 4 [$F_{(1,195)} = 8.17$, $p = .005$, $R^2 = .04$] but not at Grades 8 or 10. At Grade 4, qualifying percentage agreement decreased by about 2% for each 10 year increase in age.

Table 5 summarizes changes in qualifying status between the pre-qualifying and qualifying tests (%QStatusChange) by grade level. Pre-qualifying passing rates were comparable at Grades 4 and 8, but performance on that test was slightly lower at Grade 10. Raters at Grade 4 exhibited only a small increase in the percent passing following pseudo-scoring (less than a 2% increase), although this increase was still statistically significant. However, larger statistically significant gains in passing rates were realized by Grade 8 and Grade 10 raters (24% and 11%, respectively).

Table 5. Qualifying Status Change by Grade Level

Grade	Pre-Qualifying Passing %	Qualifying Passing %	%QStatusChange
4	47.4	48.8	1.4^{*}
8	47.8	72.0	24.2^{**}
10	27.4	38.2	10.8^{***}

Note: $N_{\text{Grade 4}} = 211$, $N_{\text{Grade 8}} = 232$, $N_{\text{Grade 12}} = 270$. * $\chi^2_{(1)} = 11.30$, $p < .0009$. ** $\chi^2_{(1)} = 10.55$, $p < .002$. *** $\chi^2_{(1)} = 12.87$, $p < .0004$.

Subsequent analyses of qualifying status gains focused on the stability of these gains across rater demographic and professional experience groups. No statistically significant differences were observed at the three grade levels for Gender, Ethnicity, Age, Teaching Experience, or Highest Degree. Analysis of rater Scoring Experience produced a statistically significant difference at Grades 8 [$\chi^2_{(1)} = 4.35$, $p = .04$], but not at Grades 4 and 10. At Grade 8, experienced raters who would have failed to qualify on the pretest were about twice as likely to qualify on the posttest than were non-experienced raters who failed to qualify on the pretest [OR = .17 / .08 = 2.00].

Scoring

Table 6 summarizes inter-rater correlations achieved during pseudo-scoring and operational scoring by grade level for raters who achieved qualifying status. Note that the values of these correlations are attenuated due to the fact that we could only compute these indices for raters who qualified on the posttest. Hence, these correlations probably underestimate the true correlations for a population of pre-qualified raters. Inter-rater correlations during pseudo-scoring were slightly lower at Grade 10 than they were at Grades 4 and 8. The increase of the average inter-rater correlation ranged from .05 to .10 across the grade levels. In addition, between 4% and 12% of the individual raters who qualified exhibited a statistically significant increase in inter-rater correlations between pseudo-scoring and operational scoring. A similar trend was observed for correlations between ratings on validity scores (i.e., between a rater's scores and expert consensus scores), but those statistics are not summarized in this report.

Table 6. Inter-rater Correlation for Qualifying Raters by Grade Level

Grade	Pseudo-Scoring Inter-rater Correlation	Scoring Inter-rater Correlation	IRCorrChange
4	.65	.73	12
8	.66	.71	4
10	.57	.67	7

Note: $N_{\text{Grade 4}} = 67$, $N_{\text{Grade 8}} = 145$, $N_{\text{Grade 12}} = 88$. Values are averaged across raters within each group.

Discussion & Conclusions

Our results suggest that the practice of pseudo-scoring may result in improvements in the performance of raters beyond what they gain from a typical training session for an operational scoring project in writing. However, these gains are modest, so decisions to engage in pseudo-scoring should consider whether the cost of pseudo-scoring justifies the increase in rater retention. In our case, we observed improvements in performance on qualifying tests (in terms of both the percentage of agreement, we observed average gains across groups ranging from about 0.1% to about 10%, and qualification rate, we observed average increases across groups ranging from 4% to 12%) as well as inter-rater agreement during operational scoring for raters who achieved the qualification criteria (we observed increases in the average interrater correlations across groups ranging from .05 to .10). At Grades 8 and 10, the increase in the percentage of agreement on qualifying sets was statistically significant, and the increase in qualifying rates was statistically significant at all three grade levels. In addition, for raters who achieved the qualifying criteria, inter-rater correlations increased at all grade levels by a reasonably large amount, and around 10% of the raters exhibited a statistically significant increase in their inter-rater correlations. Finally, we observed little evidence of a differential influence of pseudo-scoring across rater demographic and experience groups. The few statistically significant group differences that we found tended not to be replicated across grade levels.

Due to the fact that we are aware of no prior research regarding the efficacy of pseudo-scoring, we are unable to draw direct comparisons between our results and those of prior research. We should also acknowledge a few of limitations of our study. First, our study's design is not ideal. As a result, we cannot unequivocally declare that pseudo-scoring was the cause of the observed increases in rater retention and performance. Specifically, we had no control group, so, although we think these potential explanations are weak, it is possible that the observed increases were due to exposure to raters having an increased familiarity with the qualification materials due to pretesting or lower motivation to perform well on the pretest. In addition, we did not control for the amount of pseudo-scoring in which raters engaged. Rather, we only provided for a specific amount of time during which pseudo-scoring took place. As a result, due to difference in scoring speed, the amount of practice that raters gained through the pseudo-scoring process varied considerably. Second, although our sample size is large, there is a non-trivial amount of missing data on rater ethnicity (about 20%), and we needed to collapse categories for ethnicity into a dichotomy in order to have sufficient sample size for that comparison. Hence, we warn against making premature conclusions about the fact that there were no group differences observed with respect to ethnicity. Third, although our study focuses on three groups of raters (i.e., three grade levels), our analyses still focus on a fairly limited scoring context—we only consider writing within a single state. It may be that the efficacy of pseudo-scoring varies considerably in other content areas, with other types of prompts, and with other groups of raters.

Hence, it will be important to conduct additional research on the efficacy of pseudo-scoring. Future studies should adopt a research design that allows researchers to rule out alternative explanations of observed changes in rater performance (e.g., utilizing randomization or employing a control group). In addition, it would be helpful to gather additional information from raters to help determine the impact of pseudo-scoring on other outcomes (e.g., raters' perceptions of the thoroughness of training or the fairness of the qualification process). Finally, future research should be conducted to determine what durations of pseudo-scoring are most effective. It is likely that improvements in rater performance and qualification rates begin to trail off after a certain amount of practice and that future engagement in pseudo-scoring is of limited usefulness.

However, we believe that our results suggest that pseudo-scoring, as an extension of rater training in operational scoring projects, may be a useful practice. Still, there are numerous cost-benefit considerations that should be taken into account. In our study, raters spent an additional two days engaged in pseudo-scoring, essentially doubling the labor costs associated with rater training. We have no way of determining what the optimal amount of time to allocation to pseudo-scoring is from our analyses, but those who conduct rater training activities should consider whether these cost increases warrant the magnitude of improvements observed in our study. In projects in which turnaround times are short, qualifying criteria are difficult to achieve, or the availability of qualified raters is limited, pseudo-scoring may offer a cost savings due to the fact that a greater number of raters will eventually pass the qualifying criteria. We suggest also considering a two-stage qualification in which raters who achieve the qualifying standard following training be allowed to begin operational scoring immediately, while raters who fail to qualify immediately following training engage in pseudo-scoring (i.e., assigning scores that are not recorded for reporting purposes). Those raters would then take another qualifying test at the conclusion of pseudo-scoring, and those achieving the qualifying standard would begin operational scoring. Regardless, some level of pseudo-scoring may be beneficial to all raters early in the project—in our study, those who eventually achieved the qualifying criteria exhibited non-trivial increases in inter-rater correlations between pseudo-scoring and operational scoring.

References

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*, 37–64.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, *49*, 725–747.

Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*, 363–406.

Ericsson, K. A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. *Harvard Business Review*, *85*, 114–121.

Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' perceptions of on-line rater training and monitoring. *System*, *29*, 505–520.

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*, 26–43.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*, 27–33.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*, 197–223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263–287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*, 145–178.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*, 189–205.

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*, 1–21.