The Development of a Standardized Rater Training Program on Essay Scoring: A Case Study

So Young Jang
(Chung-Ang University)


The primary purposes of this study are to identify the characteristics of modeling a rater training program and to develop an efficient training model at the University of Illinois at Urbana-Champaign. For these purposes, this study proposes that a rater training program should be standardized by accomplishing innovative systematic changes that consider multiple aspects. This study utilized a modified version of Lynch’s program evaluation model (1996, 2003) to collect evidence from different sources, including data drawn from the entire evaluation process ranging from needs analysis to a feedback system based on the final product of the evaluation. Mixed methods were proposed for the data analysis. Quantitative data analysis was proposed for analyzing the surveys, and the rating corpus. Qualitative and document analysis were also essential for analyzing relevant training materials and workshop observation as well as exploring the degree of change in the perceptions of the raters. The results of this study provide educational implications for language testing. The salient value of this study is the collaboration with stakeholders in a test administration situation. Raters’ concerns and challenges were clearly identified, shared, and resolved with the practitioners.

I. INTRODUCTION

During the past 30 years, much research on rater reliability has been conducted in performance assessments (Bachman, Lynch, & Mason, 1995; Brown, 1995; Charney, 1984; Cumming, Kantor, & Powers, 2001; Erdosy, 2004; Freedman & Calfee, 1983; Lumley, & McNamara, 1995; McNamara, 1996; Shi, 2001; Shin, 2001; Shohamy, Gordon, & Kraemer, 1992; Weigle, 1994a; 1994b; 1998); however, most of these studies have focused on scoring variability as it relates to rater background, and rating procedures of the test instrument. Moreover, the studies strongly suggest that a more
systematic training program would be the best solution to remedy the low rater reliability in essay scoring. However, studies related to the topic of the training effectiveness have been conducted but one issue is that the design of the training program was not clearly identified in the studies.

This research is distinguishable from the previous research with two reasons. First, more objective standards and effective methods for designing a systematic rater training program were identified and presented via empirical research. It is true that more empirical research is necessary for answering the fundamental problems of current training programs. Several studies (Bachman, 1988; Bachman & Savignon, 1986; Salaberry, 2000) have criticized the validity and reliability of the American Council on the Teaching of Foreign Languages (ACTFL OPI) and the ACTFL OPI Tester Training Manual (hereafter ACTFL-TTM). The ACTFL-TTM was revised in 1999, but they still cannot avoid the criticism that more effort should be made to develop a theoretical model for testing and training content based on empirical data. Salaberry (2000) and Shohamy (1990) stated that the ACTFL-TTM does not reflect the essential features of theoretical and operational models to train interviewers/raters by simplifying information which the guidelines should contain. Currently, Education Testing Service (ETS) provides an on-line rater workshop program but does not clearly provide specific information on how to design or procedures on how to conduct a rater training (www.ets.org). Therefore, the characteristics of a more systematic rater training program should be identified through the standardization of the training program.

Second, some benefits of designing a more systematic training program have been found. For example, the training workshop plays a critical role in connecting the theoretical constructs of language performance with the operational constructs of the practical situation. In other words, training helps raters use the test instrument appropriately. Through the training workshop, raters should understand the test constructs, the test purposes, and the test procedures. Secondly, the training workshop should offer opportunities to monitor individual raters’ performances by having group activities and feedback sessions. An important function of a training program is to screen whether a rater is qualified.

The primary objective of this study is to develop a systematic rater training program for the ESL placement test. For this purpose, this study presents the procedures of the workshop preparation and evaluates the effectiveness of a proposed new training program.
II. LITERATURE REVIEW


First of all, program evaluation theory provides a systematic framework for the overall research procedure as well as a philosophical background. It can be adopted in order to see the overall logic of the evaluation process of the training program. In program evaluation theory, the evaluation is used not only to improve program effectiveness but also to contribute to better policy-making based on evaluation findings. The role of an evaluator is to assist stakeholders involved in the educational program to better understand what is happening, help find shortcomings of a program and help users select appropriate information for improving programs (Patton, 1996; Weiss, 1998).

Particularly, Lynch’s program evaluation theory as a context adaptive model (CAM) in applied linguistics provides guidelines about what should be evaluated and how we should approach the evaluation of each phase (Lynch, 1996, 2003). Lynch has discussed the context adaptive model for program evaluation as it is applied to language education and testing. This model proposes that the entire evaluation procedure should reflect the specific concerns of stakeholders and should be designed depending on the context of the particular program being evaluated.

In line with this perspective, program evaluation procedures could be considered a possible evaluation model for realizing the standardization of a rater training program by covering the entire scope of evaluation. Lynch’s model is context adaptive and provides seven phases in carrying out the program evaluation. The model is responsive to the evaluation environment by providing for evaluation research phases from analysis of the audience and elaboration of the evaluation goals as a result of two-way communication with stakeholders and the evaluation context.

There are two critical issues to be discussed. One is the determination of how evaluation research is characterized--for instance, formatively or summative. Second, the primary attitude toward how to design the research should be decided. Which approach is more suitable for evaluation research--positivistic approach, interpretive approach, or mixed approach? This is an important decision in the sense that these two efforts help determine specific methods for data collection and analysis.

Evaluation can be either formative or summative, depending on the goals of the evaluation. Formative evaluation is seen as an assessment of the ongoing progress and demands of the participants in a particular program while that program is developing. It has instructional purposes that focus on what individuals have achieved and how
well the program is functioning. The final product of formative evaluation suggests some changes for improving the quality of individual performance and the program. The concern of summative evaluation, on the other hand, is high-stakes decisions about the worth of a program—whether the program can achieve its goals and expectations, and whether it is worth financial support.

Lynch (1996, 2003) has discussed a positivistic perspective as a paradigm of program evaluation in detail, since this approach plays an important role in determining research design. First, the positivistic perspective is concerned with two things: certainty and generalization. Positivists are interested in accurate measurement and inferences from the program, and they try to identify the relationship between causes and effects. They are also interested in generalizing the evaluation results to other evaluation settings. This approach suggests quantitative data gathering during short or long term periods. A certain format of treatment is provided for a program group to estimate the effectiveness of treatment. This approach tends to prefer the use of experimental or quasi-experimental design, and the outcomes of the program group and comparison group are compared to see the impacts of treatment. Research design can be determined depending on the evaluation context. For this study, formative evaluation employing quasi-experimental research design was conducted to estimate the effectiveness of the program after a training workshop and curriculum development for raters.

2. Standardization of a Rater Training Program

Standardization based on Training theory, originally adopted from the human resources field, provides a link between the theoretical issues and practical concerns. Because of uncertainty about the format of a test instrument and the stakeholders involved in the rating environment, training programs, both onsite and online, are not always successful at enhancing test validity and reliability (Elder, Barkhuizen, Knoch, & Randow, 2008). The goal of standardization is to institute an entire rating system, including rater training, and to promote rater professionalism with institutional support, rather than assigning the responsibility on low score reliability to individual raters (Fulcher & Davidson, 2007). It seems that an effective training program could be achieved through standardization because standardization can provide an equal quality of training materials and practices. On the basis of training theory analysis, the most appropriate training materials and methods should be created in cooperation with the trainer and staff of a language program.
For the development of a systematic\(^1\) training model to fit the rating context, theoretical training models, specific goals, and methods of the training program should be defined for a particular training program. First, the systematic training program should make the rating focus clear by presenting and sharing accurate information (Borman, 1977, 1978, 1979). A standardized rater training program adjusts the balance between knowledge and experience of raters. Recent research (Brown, 1995; Choi, 2000, 2002; McNamara, 1996; Saal, Downey, & Lahey, 1980; Shin, 2001; Steward, 1999; Weigle, 1994a, 1994b) has consistently shown that rater training plays a role in reducing ambiguity and maximizing shared knowledge among the raters, although there exist individual differences in rating schema, background knowledge, and experience in essay rating. For instance, during a training workshop, raters communicate with each other and with the trainer to increase shared knowledge and decrease ambiguity through the allocated rating tasks. Increasing shared knowledge and rating schema, therefore, helps raters comprehend rating tasks and facilitates the understanding of the intended test purpose or rating purpose. Through a standardized training program, sufficient information on the rating process itself and the rating context are necessary for avoiding or reducing rater bias. This provision can in turn reduce individual differences. To enhance the effectiveness of a training program, a more systematic and formulated training program is necessary.

Second, Bernardin and Buckley (1981) discussed the fact that training effectiveness should be evaluated based on multiple sources in terms of training design (modeling), training methods, measures of rating error, rating scale descriptors, and the nature of rating tasks including the characteristics of the examinees. In addition, these features, which might affect the procedures of the rating, should be standardized through an appropriate consensus process. Several standards for measures of rating skills should be identified (Hauenstein & Foti, 1989; Lunz, Wright, & Linacre, 1990; McNamara, 1996; Roch & O’Sullivan, 2003; Shin, 2001; Weigle, 1994a, 1998). In addition to this, measures at the program level should also be considered on the ground of empirical findings. In order to maximize the effectiveness of the training program, it might be necessary to take into account these substantial features which might affect the quality of the assessors’ performance.

---

\(^1\) The terms “standardization” and “systematic” are conceptually almost interchangeable in this study, but the two terms can be distinguishable. “Standardization” can be defined as the series of activities or procedures which make a rater training program more systematic. A “systematic” training program can be defined as the final product of standardization. In this study, these two terms were used interchangeably.
III. RESEARCH DESIGN

1. The ESL Placement Test (EPT)

The data for this study were collected from the ESL Placement Test (EPT) at the University of Illinois at Urbana Champaign (UIUC). The EPT is a test for placing new international students into the appropriate levels of ESL classes (http://www.linguistics.illinois.edu/students/placement). It measures two different components of language ability: speaking and writing but this study, however, focuses on the rating process of the essay test.

The essay test is a paper-based test which takes almost three hours to complete. Test takers have consecutive activities with an ESL teacher and peers, who are sitting next to one another. At the beginning of the test, the EPT procedures and topic are introduced, and test takers are provided with a topic-relevant article. After reading the two-page article, the test takers participate in a 30 minute mini-lecture and discussion with an ESL teacher and peers.

There are three topics for the EPT test: globalization, cloning, and animal testing. One of the three topics is randomly chosen, and an ESL teacher gives a short lecture session defining the basic terms using an overhead projector, and giving directions about how to collaborate with peers. The teacher also briefly explains the scoring guidelines to the test takers. The test takers make a first draft of their essays, and then they review it with their peers. Finally, the test takers individually write a final draft of their essays. In this stage, the test takers have 60 minutes to write a two- to three-page essay.

2. Participants

1) The Raters

The recruitment procedures were carried out in two phases, and the raters and coordinator of the ESL writing courses (trainer) participated in this study. Three different recruiting processes were conducted to meet the research requirements, and participation in each research phase was voluntary. In particular, for the post-rating, six raters were recruited and three who participated in the proposed new workshop program were assigned to the workshop group. This means that three raters served as ESL teachers until this research was completed. However, control group raters conducted essay scorings without the training. They have ESL teaching experience, but did not serve as ESL teachers while the research was underway.
2) The EPT Trainer

The coordinator of the ESL writing courses (EPT trainer) had a lot of ESL teaching and rating experience. She was actively involved in designing and implementing the new training program.

3) The Role of Researcher as Evaluator

As an evaluator, the researcher conducted this research in collaboration with the EPT trainer. The researcher has explored the demands of the EPT raters and the trainer through survey and in-person meetings. The researcher has a meeting with the trainer to discuss the current issues based on the results of needs analysis and share ideas for the workshop preparation.

3. Research Questions

Research questions were formulated for each stage of the proposed new model. This study was guided by the following questions:

1. What issues about the evaluation context do stakeholders perceive?
2. Which part(s) of the training program can be standardized? How can the EPT training program be standardized?
3. Was the proposed new training workshop successfully conducted?
4. How effective is the proposed rater training program?

4. Rationale for Data Collection and Analysis Using Program Evaluation

Table 1 exhibits the distinguishable features of the revised Lynch’s model used, the research design and specific research activities corresponding to major research steps in this study. Lynch’s model was adopted and modified for this study. The research phases of the revised Lynch’s context adaptive model (CAM) are also related to the evaluation model.

Lynch’s context adaptive model (CAM) originally proposed seven major steps: step 1 is Audience and Goals, step 2 refers to Context Inventory, which includes needs analysis and step 3 is Preliminary Thematic Framework, which is a stage of literature review in order to set up solid theory to fit any context. Steps 4, 5 and 6 refer to Data Collection Design, Data Collection and Data Analysis. Finally, step 7 is an Evaluation Report, which is a stage to make a final research report.

The revised model maintains the seven major phases, but the research activities at
steps 4, 5, and 7 were revised to accommodate conducting a program evaluation of the rater workshop, since that requires more interaction with stakeholders. Interim outcomes from each phase are necessary. In Lynch’s model, step 4 was originally designed for data collection, but in this study, it was eliminated, after careful consideration. Step 4 became “Standardization of a rater training workshop.” Step 5 became “Implementation of training workshop,” and step 6 & 7 were changed as post-rating and data analysis, but is the almost same as Lynch’s original model. Regarding the evaluation steps, four research questions were formulated; Table 1 describes the purpose of each research question.

In order to answer the four research questions, the data for this study were collected from the ESL Placement Test (EPT) at the University of Illinois at Urbana-Champaign (UIUC). The data were collected in the steps from July, 2009 to January, 2010. The target audience of this study was comprised of the people involved in the EPT essay test. Different participants were involved in the different phases of this research. They were all recruited from the same population, because they were all ESL teacher raters.

Employment of mixed methods was appropriate for this study because data were collected from various sources such as the pre-workshop survey, review of internal documents, meeting notes, the post-workshop survey, and results of the post-rating during the different research steps. As a result, it was necessary to use mixed methods when analyzing these data. The data were analyzed using descriptive statistics for analysis of surveys, and FACETS and GENOVA were used to measure the raters’ performance in the post-rating. As a qualitative analysis, document analysis and content analysis of the training materials were adopted, and meeting notes and workshop observation were analyzed.

Training effectiveness was estimated at several layers: Interim outcomes of each stage were analyzed and reported at the individual level and some changes were evident at the program level. In addition, mixed methods yield a more meaningful interpretation, rather than a biased interpretation which might have arisen from using a single method, due to limited findings.
TABLE 1
Overview of the Research Design

<table>
<thead>
<tr>
<th>Step</th>
<th>Question</th>
<th>Purpose</th>
<th>Data Source</th>
<th>Participants</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &amp; 2</td>
<td>Q1: Pre-workshop analysis I</td>
<td>Understanding the evaluation context and identifying the issues or concerns of stakeholders</td>
<td>Pre-workshop survey Internal EPT documents</td>
<td>Eight raters and the EPT trainer</td>
<td>Descriptive statistics and document analysis</td>
</tr>
<tr>
<td>3 &amp; 4</td>
<td>Q2: Pre-workshop analysis II</td>
<td>Identifying the appropriateness of the proposed new training workshop</td>
<td>Meeting notes</td>
<td>Researcher collaborated with trainer to elaborate training materials</td>
<td>Content analysis to modify the training materials</td>
</tr>
<tr>
<td>5</td>
<td>Q3: Workshop stage</td>
<td>Identifying successful implementation of the training workshop</td>
<td>Observation of the workshop</td>
<td>Fifteen raters who were as ESL teachers</td>
<td>Observation notes and descriptive statistics for the surveys</td>
</tr>
<tr>
<td>6 &amp; 7</td>
<td>Q4: Post-workshop analysis I</td>
<td>Evaluating the effectiveness of the proposed EPT training program</td>
<td>Post-workshop survey Post-rating data</td>
<td>Six raters</td>
<td>Descriptive statistics, FACETS &amp; GENOVA</td>
</tr>
</tbody>
</table>

IV. RESULTS

For this study, the data collected were analyzed and presented on the basis of three major research stages: analyses of pre-workshop, workshop observation, and post-workshop. Pre-workshop analysis was conducted with the data which were collected at evaluation steps 1, 2, 3 and 4. The data for workshop observation were collected from evaluation step 5. Finally, in the post-workshop session, quasi-experimental research design and post-workshop survey were employed to estimate the effectiveness of the program by comparing the two rater groups after a training workshop.
1. Data Collection, Analysis, and Findings for Question 1:
Understanding the Evaluation Context and Identifying the Issues or Concerns of Stakeholders

1) Data Collection and Analysis of Step 1 and 2

A needs analysis at evaluation step 1 and 2 was necessary to understand the program context. A deeper understanding of the context of a particular program is important to program evaluation. Needs analysis based on document analysis and surveys provided knowledge of how well the program worked; what issues need to be identified; and what the beliefs or expectations of the people involved in the program were, regarding the effectiveness of the training program. There were three different sources of data: the current training materials, the test specifications, and a survey of the stakeholders.

Documents were collected from the website (https://netfiles.uiuc.edu/fgd/www/ept_bulletin.pdf), and some materials for internal use provided by the coordinator of the ESL writing courses and EPT administrator were analyzed. The EPT bulletin provided information about how to register and prepare for the test. It also included test procedures, descriptions of test tasks, and ESL course guidelines. Scoring guidelines, provided by the EPT administrator, were critically reviewed in order to revise the training program. In addition, the current EPT training program was analyzed based on documents analysis of training materials.

The surveys were administered to the raters and the trainer. The survey consisted of two parts: questions using four-point Likert scales, and open-ended questions. For raters, the questions covered various topics related to the EPT test, scoring procedures, the current training program, and trainer evaluation. For the trainer, some open-ended questions about designing a new training program were added. Eight ESL teacher-raters who were teaching or had taught ESL writing classes participated in this survey, and the survey was collected in person or via e-mail at the participant’s convenience. Because of the small number of participants, descriptive statistics were used in analyzing survey questions, and the responses to the open ended questions were categorized based on themes/topics and analyzed.

2) Setting-up Evaluation Goals

The primary objectives of this study were to evaluate and develop a systematic rater training program for the EPT with a fuller perspective of rater reliability. A secondary goal was to identify the effectiveness of the proposed training workshop based on the
stakeholders’ perceptions and measures of their performance.

3) Understanding the Operational EPT Rating Environment

After the essay test, the EPT administrator distributes the essay the same afternoon, and normally two raters participate in the marking process. Rating essays is one of the responsibilities of the ESL teachers, assigned on their duty rosters. Two raters are asked to give a single holistic score for each essay. The holistic scores consist of four levels which are related to the ESL writing courses (Too low, ESL 113/500, ESL 114/501, and ESL 115/Exempt). Next, if there is a discrepancy between the scores assigned by the two raters, discussion is required to reach an agreement. If the two raters do not reach a consensus through discussion, then a third rater decides the score.

4) Findings from the Review of the Training Materials

The ESL course guidelines for internal use, including the prototype essays, were collected and reviewed. In addition, all of the scoring materials provided by the EPT administrator (a graduate assistant) were critically reviewed in order to standardize the training program. The ESL course guidelines provided the EPT benchmarks for both graduate and undergraduate levels along with the different level of ESL writing courses. The holistic benchmarks described each proficiency level and simple directions for connecting essay rating to ESL courses.

The EPT trainer had some old prototype essays and answer keys for rater training. Twelve essays were packaged as a practice set, and an answer key provided the correct proficiency level for each essay, as well as a short justification for the assigned score. The EPT bulletin provided general issues of the essay test for examinees, including the purpose of the test, test preparation, test process, scoring, and ESL course registration. The EPT administrators are given the most current version of the test prompts/directions, with test topics and lecture notes that proctors have used in testing situations.

The 12 prototype essays that the EPT had been using were reviewed and recommended eliminating some misplaced essays modifying others, which the researcher and trainer discussed. The researcher and trainer agreed that a new selection of prototype essays was essential, because the existing essays were out of date and did not represent the current version of the EPT test. The answer keys were removed, because face to face contextualized feedback was considered more helpful. Standards for essay selection in terms of the number of essays and test topics were suggested.
5) Findings of Pre-workshop Survey from the Trainer

The EPT trainer was asked to fill out a pre-workshop survey via e-mail for her convenience. The survey asked open-ended questions about two topics: the EPT training workshop and workshop preparation/management. The trainer responded that the role of the EPT trainer was to lead the workshop and serve as a fourth reader, providing some justification for ESL placement when disagreement occurred in the rating. Second, she responded that the primary purpose of the EPT rater training is to re-familiarize T.A.s with the EPT essay tests and benchmarks, and to calibrate raters to the benchmarks. Finally, the trainer pointed out that a lack of instructional resources for raters’ professional improvement was a big challenge for her. She specifically suggested that instructional materials be standardized and revised to implement a well-organized workshop.

The trainer’s survey asked about plans to organize and manage the rater training workshop. The trainer provided a general idea about the rater training program based on her rating experience, saying that the workshop program had been designed both to familiarize raters with the EPT rating system and to provide rating practice with prototype essays. Regarding to the focus of the EPT training program, she considered rater reliability and rating accuracy as more important than agreement among raters. She also mentioned her concern that a 2-3 hour workshop and the limited number of prototype essays were not sufficient to represent the rating issues that EPT raters may encounter. These issues were discussed, and some of the content was modified when preparing the new training program.

6) Findings of Pre-workshop Survey from the Raters

The results of the pre-workshop survey showed that raters were satisfied with the EPT rating system overall (double rating, holistic rating, and consensus process are allowed); nevertheless, the findings suggested that several area should be improved. First, the findings suggested that the holistic rating scale descriptors needed to be improved. The accuracy of the rating scale descriptors and assessment criteria needed to be reviewed. Particularly, rating scale descriptors at the higher levels were considered relatively less accurate. In addition, rating scale descriptors needed to be evaluated and modified through open discussion with raters.

Second, in terms of test topic and test procedure, it was found that raters showed a positive attitude toward the relevance of the test topics to the measurement of writing ability. However, raters were concerned about the extent of the students’ active participation in the test procedures and disadvantages of the test topics for individual
students. These findings suggested that individual raters have different perceptions of the difficulty level of the topics, which showed that the difficulty of the topics might affect the rating context. With respect to assessment criteria, some raters felt confused about content, grammar and lexical choice, and use of sources; however, plagiarism and organization had clearer criteria to raters.

Interestingly, although an analytic rating system was not used at the time, the raters felt such a system might be useful for reaching a more accurate decision in a practical situation. Regarding the consensus process, the two raters with the least rating experience answered that they were likely to change their final decisions after discussion with peers. This finding showed that teaching experience was a salient feature when justifying their rating decisions.

In past workshops, the EPT training had focused on rater agreement, rather than on rater reliability (consistency) and rating accuracy. In the new workshop, rating accuracy and consistency needed to be strengthened by revising or creating new rating materials. Next, raters showed enough self-assurance about rating, whereas half also wanted further training. More iterative feedback was suggested for the new workshop, and raters wanted to have more practice rating.

Finally, some problems were clearly identified in terms of the utility of the workshop rating resources. There was no independent EPT workshop package for essay rating, and the training materials were not organized. ESL coursework guidelines provided the only holistic descriptors, and essay rating was not a main issue of the T.A. workshop. In the EPT bulletin, the focus of the information was the description of the EPT test procedures, rather than essay rating. It was suggested that the information needed to be updated and revised. Some of the materials had already been updated, but some were out of date. It was pointed out that the prototype essays for practice and their answer keys should be updated, because they no longer reflected the current essay topics and rating issues.

Increasing accessibility to training materials by synthesizing and reorganizing all relevant materials was an important suggestion. The trainer and the EPT administrator had separate information, and it seemed that they had no opportunity to actively share the rating materials, although the EPT administrator had access to more information, such as test procedures, test topics, test prompts/directions, and some materials used in the operationalized EPT test. It seemed that the raters had fewer chances to see the rating materials.
2. Data Collection, Analysis, and Findings for Question 2: Identifying the Appropriateness of the Proposed New Training Workshop

1) Data Collection and Analysis of Step 3 and 4

On the basis of the analysis of the audience, the evaluation goals, and the evaluation context, some evaluation arguments and critical issues arising in the current program can be provided in the preliminary theoretical framework. The framework makes the focus of the evaluation clear and assists the evaluator in carrying out research design, data collection and analysis. Detailed content materials, activities, training methods and institutional expectations in terms of reliability (agreement/consistency) were determined based on a review of the literature surrounding training theory and language testing theory.

In Step 3 (preliminary thematic framework), based on the interim outcomes of Step 2, the researcher confirmed the findings of previous studies so that an appropriate workshop framework could be formulated.

With respect to evaluation step 4 (preparation of the EPT training workshop), in consideration of the theoretical paradigm and the practical demands, a discussion with the EPT trainer about how to achieve standardization of the rater training program, including training input and activities and institutional support, was necessary. The literature review and the data collected at Steps 1 and 2 were used to revise the existing workshop materials. To prepare the workshop, the researcher met with the EPT trainer and collaborated on developing the training workshop, revising the EPT training curriculum and workshop plans as needed. The researcher shared her knowledge about the literature surrounding rater training with the trainer by providing a short summary of the findings of previous studies and the results of pre-rating. In collaboration with the trainer, the researcher decided the workshop schedule and how to standardize the workshop materials and procedures.

Data were collected from two sources: my notes from meeting with the trainer and the training materials. The researcher had biweekly meetings with the trainer from October, 2009 to December, 2009, and three more meetings in January, 2010, before the workshop. The researcher had made field notes to summarize what aspects of the workshop curriculum needed to be changed, with rationales. After discussion with the trainer, some changes were recorded in the form of field notes. On the basis of the results of the discussion, the researcher provided a summary of the rationale for replacing training content, training methods, and prototype essays. This included a summary of the findings of the pre-survey and some suggestions for revisions based on the literature.
2) Findings of Meeting Notes

(1) Meeting 1

The researcher had the first meeting with the trainer on November 11th, and explained the purposes and schedule of the research. The pre-workshop survey was given to the trainer, basic information was shared, such as the characteristics of the rater group, the nature of the training materials, and the workshop schedule. It was decided that the trainer would review the current training materials and bring some ideas to the next meeting about how to organize the schedule what kinds of training materials to include. In addition, it was agreed that the researcher would provide the results of the pre-workshop survey administered to raters. It took approximately 30 minutes to cover the agenda for the first meeting, and the next meeting was scheduled for November 20th.

(2) Meeting 2

The main point was discussed at the second meeting to explore the weaknesses of the workshop program. First of all, the pre-workshop survey results and raters' concerns based on their responses to the open-ended questions were reported: the raters wanted a more organized workshop program with accurate rating scale descriptors, more practice, and more contextualized feedback. Particularly, raters had some difficulty deciding about borderline essays, and they wanted rating principles/tips. The theory from the literature review and the practical findings of the needs analysis from Steps 1 and 2 were used as the basis for revising the existing workshop materials.

(3) Meeting 3

Again, rough workshop plans and activities were discussed. First, it was agreed to gather all relevant training materials. Electronic documents containing test directions for examinees, three different test topics, and analytic guidelines were obtained from the EPT G.A. The trainer provided holistic guidelines for the undergraduate and graduate levels, and 12 prototype essay samples used in previous workshops. Finally, the workshop schedule was discussed. The trainer provided a rough idea of the workshop program and wanted to elaborate on her specific plans at the next meeting. The following is the basic outline suggested during this meeting for the workshop program: (a) review of sequence and goals, (b) overview of benchmarks, (c) individual rating & whole group comparison 1 (3 essays), (d) individual rating &whole group comparison 2 (3 essays), (e) discussion of rating procedures (what to do with borderline essays, and what to do when 2 raters disagree), and (f) small group
rating and comparison and discussion of trainer feedback.

Employing analytic scoring and selecting new prototype essays were the major issues discussed at this meeting. The researcher proposed that adapting analytic guidelines might be helpful to guide the raters’ decision-making processes for borderline or inconsistent essays, because almost half of the raters were new raters with only one semester of teaching experience. It was decided that analytic scoring methods would be employed for the workshop, and an old version of analytic scoring guidelines was modified by the trainer. Regarding to prototype sampling, the trainer wanted to choose new essay samples and revise the answer keys, because they were out of date. The trainer provided the old prototype samples and answer keys for the researcher. In addition, the trainer asked about some issues to be discussed at the next meeting: the principles or standards for prototype essay sampling, and the number of essays to be used for the workshop.

(4) Meeting 4

In early January, the researcher and trainer met again to decide how to conduct the workshop program with the revised training materials, and to confirm the specific workshop schedule. The trainer provided a workshop plan including dates, times, training focus and plans. The workshop was scheduled for January 12th, and the computer lab was reserved for three hours. The researcher presented a short report for the trainer based on three points, and the trainer’s feedback was positive. The researcher double-checked all research plans with the trainer. The researcher asked for some time to have the raters fill out the consent letter before the workshop started, and for permission to audio-record the entire workshop session. Finally, the researcher asked to collect all of the rating sheets from the activities, and at the end of the workshop, conduct a survey about raters’ evaluation of the workshop program. It was agreed that the workshop materials and activities would be developed and modified as the researcher had suggested, but some parts were not accepted due to practical reasons, such as time constraints or fatigue. The trainer developed Power Point slides for the workshop and prepared some activities with the new prototype samples. In addition, the trainer wanted to develop new analytic scoring guidelines and rating sheets for the practice session, rather than modifying the old guidelines. It was decided to have one more meeting before the workshop in order to confirm the workshop materials.

(5) Meeting 5

The fifth meeting was scheduled for January 8th to finalize the workshop procedures and activities. The appropriateness of the scale descriptors, assessment
criteria, and scoring methods was evaluated and revised to provide a solution to individual raters’ problems. Some parts of the training materials had already been already well prepared, but some had not been because of time constraints.

(6) Meeting 6

The researcher met with the trainer again to double-check the workshop program. This was the last meeting, held on January 11th, one day before the workshop. The trainer shared a website to which the modified training materials had been uploaded, the UIUC ESL Writing TA home. In addition, the trainer and researcher confirmed the holistic and analytic scoring guidelines, feedback worksheet for obtaining feedback about the revised rubrics from raters, and the rating sheets used for the workshop activities.

3) Characteristics of New Rater Training Program

(1) Clear Training Focus and Adopting More Interactive Methods

The organization of the workshop program and the delivery method of the information were improved. Training materials were incorporated and reorganized to fit the EPT rating context. For instance, based on the analysis of raters’ evaluation on the old rater training program, the lecture session was not very helpful because the training focus was not unclear. Besides, raters want to practice more essay samples. Both individual and group activities were employed to improve rating skills, including feedback from peers and the trainer.

(2) Selection of Representative Prototype Essays

In the new program, new instructional resources should be created and developed to strengthen the current training program. Another problem of the old training program is that essay samples were out of date and they suggested replacing them with new samples. New prototype essays were selected for practice during the workshop. The prototype essays used in the workshop were selected by the EPT trainer. Representative essay samples for varying proficiency levels were selected and used in the practice and discussion sessions of the training workshop in order to enhance raters’ understanding of making a scoring judgment. Prototype essay samples were obtained from the EPT data bank. Twenty-nine essay samples at both undergraduate and graduate levels were selected, including prototypes of each proficiency level and borderline essays. The three test topics were almost equally represented. Ten essays each were written for cloning and animal testing and eight essays were written about globalization. These selections were made based on the results of the document
analysis.

(3) Analytic Scoring Guidelines and Rating Sheets

The EPT trainer developed new analytic scoring guidelines with a 12-point rating scale. An analytic scoring method was adopted for the purpose of rating practice since analytic guidelines provide more accurate information for raters who lack teaching experience. In addition, five new assessment criteria were developed by modifying the previous training materials: focus, support/elaboration, organization, conventions, and integration. A rating practice sheet was provided for raters to record both analytic and holistic scores, their opinions, and group activities.

(4) Creating an Interface with Web Tools

In the new proposed training program, the workshop program was more organized, and the training materials were carefully modified to fit the needs of raters and were updated on the T.A. website. All newly developed training materials were uploaded to the UIUC ESL Writing TA home page (http://uiuceslta.blogspot.com/), including the Power Point slides used in the workshop, the electronic documents about the EPT test procedures, the test directions, test topics, the holistic and analytic scoring guidelines, the rubric revision worksheet for obtaining feedback from raters, and the workshop rating sheet. Using a blog is efficient and interactive. EPT raters can access the materials at their convenience without restrictions of time or place. Moreover, raters can leave comments or questions about the rating activities, share opinions with peer raters, and get feedback from the trainer. This could be an ongoing process throughout the semester, and is one way in which the training program could be enhanced for the next workshop.

3. Data Collection, Analysis and Findings for Question 3: Identifying Successful Implementation of the Training Workshop

This section provided evidence about how well the workshop plans had been implemented with respect to evaluation stage 5. The trainer provided the initial idea for the new workshop program, and the researcher fleshed out the idea.

1) Data Collection and Analysis

There were two data sources: observation of the workshop and the materials, and products of the workshop. The researcher observed and took notes on the entire two workshop sessions and the three activities and the consensus process were audio-
recorded. For rating practices, all materials and products of the activities were collected and analyzed.

2) Findings of Standardization of the Entire Workshop Program

Through rigorous discussions between the researcher and the trainer based on interim findings from Steps 1, 2, and 3, a clear picture of the workshop was presented, and the training workshop was conducted by the EPT trainer. The workshop content would be presented with Power Point slides; implementation of the practice sessions was discussed. The holistic guidelines were evaluated and the current descriptors were kept analytic scoring guidelines and new assessment criteria were developed. Rater feedback on the criteria and descriptors would be solicited after the workshop. New rating sheets were created for the rating practice. Both experienced and new raters participated in the training. The training workshop was expected to last approximately three to four hours. The training program provided a short familiarization and norming session to understand the rating context. In addition, raters were asked to complete a survey to evaluate the revised training program.

(1) Phase 1: Familiarization

The rater training program was modified based on the current content and workshop program of the EPT essay test. The training program was designed to provide short familiarization and norming sessions to understand the EPT rating system. During familiarization, general information on the EPT test and rating procedures would be provided. First, in the lecture session, knowledge about several rating topics as well as general information would be delivered. The trainer would define the general rating guidelines and provide the raters with information about the purpose of EPT rating, holistic guidelines, test topics, and scoring procedures.

(2) Focus of Activity 1

The purpose of Activity 1 is to familiarize raters with the EPT rating system. Three prototype essays were distributed to raters, and the holistic scores were disclosed. The raters provided analytic scores based on the holistic scores. The essays selected for Activity 1 represented the typical characteristics of the three different proficiency levels and were easy to rate. The five analytic scores for these essays supported the holistic score by clearly matching the general features of the essays and scale descriptors. The final scores were provided for raters; Activity 1 was designed so that raters could familiarize themselves with the general features of each proficiency level without guessing. For this activity, the raters were divided into five pairs. If they had
difficulty determining a score, they looked at the rating scale again and made the best
judgment they could, or they made a note of the reasons for their difficulty. The
follow-up session allowed them to discuss the issues they had encountered.

(2) Findings of Activity 1

The five analytic scores for each essay were added together and converted to a
holistic score based on the score range. Raters assigned a holistic score based on
global essay evaluation. Four raters' scores across three essays perfectly matched the
trainer's scores, but five raters' on Essay 3 were different from the trainer's score. R9
showed some confusion deciding between 114 and 115 as a holistic score on Essay 2,
and R12 felt confused between 114 and 115 on Essay 1. It was found that for Essay 3,
a holistic score based on global impression was more accurate than the analytic score
in comparison with the original and trainer's scores. The findings suggest that the
analytic scoring method should be improved to avoid score lowering.

(3) Findings of Group Rating

Raters had one more chance to revise their scores through a discussion with peer
raters. For example, in rating Essay 1, all five groups' scores perfectly matched the
trainer's score. For Essay 2, four groups reached a consensus on the ESL 115, but
Group 5 considered it to be on the borderline between ESL 114 and 115. Regarding
Essay 3, four groups agreed on ESL 113, but Group 1 did reach a consensus. Group 6
in the second workshop showed that the rest of the raters of the second workshop
group agreed with the original score after group discussion. The findings showed that
group consensus may be helpful for reaching a standard score.

(2) Phase 2: Norming Session: Individual and Group Rating.

The focus of the calibration session was how raters apply their understanding of the
basic principles of rating. Accurate observation and selection of appropriate
information from the essay were emphasized. Raters obtained feedback from the
trainer about how to match essays with scale descriptors for accurate categorization.
Through the prototype sample practice, raters could experience evidence-based
decision making.

In the norming session, individual and group practices were implemented, and
trainer feedback also was provided. Raters assigned analytic scores first, then holistic
scores. Otherwise, the rating procedure was the same as Activity 1. After finishing the
individual rating, raters went through a group consensus process, comparing their
individual scores and sharing their experiences and opinions. After the rating, the
trainer confirmed the scores of the prototypes with the class, and discussed rating
issues again. They discussed holistic or analytic scores with the trainer when a discrepancy arose with the standard scores. This process was repeated until all essay samples were completed. All raters were involved in the group discussion session to share and resolve their concerns.

① Training Focus of Activity 2

The purpose of Activity 2 was to calibrate raters’ decision making process. Raters were re-grouped, into groups of five, and each group had the same eight essays. This activity was different from Activity 1 because the individual decision making process was considered more important. No information was provided about the essays, and individual raters spent more time on individual decision making.

For Activity 2, eight essays were selected from a pool containing all proficiency levels and both undergraduate and graduate essays. Essays were selected for Activity 2 with the aim of training raters to rate borderline essays. This group of essays included complex features, reflecting inconsistent performance across assessment criteria. These essays required a higher level of cognitive demand or rating strategy, because they did not perfectly match the scale descriptors.

② Training Focus of Activity 3

The purpose of Activity 3 was to simulate and understand the basic principles of the EPT rating situation. As in the EPT rating context, individual raters assigned a holistic score and pairs of raters underwent a consensus process. If there was a discrepancy in scores, they had a group discussion session to discuss the validity of their individual scores. They also shared rating experiences and teaching experiences. After finishing individual and group rating, the trainer confirmed the scores of the prototypes and, if necessary, they discussed rating issues again. For Activity 3, the analytic scoring was excluded. The raters were assigned to groups of five and each group had two-three essays.

Activity 3 was designed to reflect the experience of rating for the operationalized EPT, essays were randomly selected. The test topic was selected depending on the rating split rate. For practice, the seventeen remaining essays were assigned for Activity 3, including four undergraduate essays at the 115 level, and six graduate essays at the 500 level. An answer key was provided during the feedback session, describing the exact score given by the trainer and the reasoning process for reaching that score.

③ Findings of Group Consensus Process from Activity 2 and 3

The summary of the consensus process yielded some interesting findings. Raters
who presented specific evidence based on the essays and rubrics tended to lead the
conversation, and consensus was easily reached. Raters who did not fully understand
the meaning of the scale descriptors had trouble rating. In addition, rating experience
may have affected the evidence-based decision making process. The more experienced
raters tended to present more evidence to support their scores.

(3) Closing and Findings of Post-workshop Evaluation

Right after the training, a post-workshop survey was provided for raters in order to
explore the degree of their comprehension of the training session and to evaluate its
usefulness. A survey was provided for raters at the end of the workshop. The survey
asked raters to mark how much the workshop had helped refresh their memories and
resolve their problems. The survey took approximately 20 minutes, and it was
collected immediately after the workshop program. Fifteen raters who participated in
the proposed training program responded the survey of workshop evaluation.
Descriptive statistics were employed for this analysis, and responses to open-ended
questions were categorized and summarized.

The findings indicate that raters were overall satisfied with the new workshop
program and its goals, schedule, content and delivery methods, and activities but that
a three hour workshop was insufficient. All raters answered that the lecture and
individual practice sessions were helpful to improve their rating. All of the raters
thought that the prototype essay samples used in the workshop were appropriate.
These findings suggest that the workshop content was helpful for raters, but that they
might still want to learn practical rating skills for the operational EPT situation. These
findings indicate that the workshop had a positive effect on raters’ perceptions about
their learning process.

4. Data Collection, Analysis and Findings for Question 4: Evaluating the
Effectiveness of the Proposed EPT Training Program

1) Data Collection and Analysis

The post-rating session was designed to investigate training effectiveness by
estimating the raters’ scoring performance. The post-rating session was conducted one
week after the workshop, from January 18th to 30th, 2010. A simple quasi-
experimental design (workshop group and control group) was proposed for this
research in order to evaluate the effectiveness of the training program (Wexley &
Latham, 2002). Originally, experimental design was proposed for this study, but it
was inevitably decided that comparisons between the workshop and the control groups
were made due to difficulty of recruitment.

In this proposed design, one rater group was the workshop group with workshop inputs, and participated in the onsite rater training session. The control group did not participate in the workshop (no input). One week after the workshop, the same essay package was given to the two groups, and the same rating procedures were followed. The 90 writing samples for the post-rating session were obtained from the EPT bank of approved writing samples, in which names and their school IDs are de-identified. The three topics were evenly chosen and the sample package from various proficiency levels was collected from the EPT essay test to explore the accuracy of the raters’ scoring decisions.

The post rating session lasted two weeks. Six raters voluntarily participated in the post rating, and three raters were assigned to each group. Three of them participated in the spring workshop program, but the rest did not. They all have ESL teaching experience. The scoring reliability of the six raters was estimated using a quantitative approach and the results of the measures were compared to see how effective the training workshop had been. Raters were asked to score selected essays in terms of both holistic and analytic scores. Rating guidelines were designed and provided by the researcher. Scoring results were collected by e-mail. The analysis focused on rating accuracy and reliability. Quantitative methodology was used in analyzing the rating data, including descriptive statistics to reveal rating patterns, G-theory and FACETS to look at rater reliability.

Finally, at stage 7 (overall evaluation), based on an understanding of overall evaluation, the different kinds of data collected from the interim outcomes at each research phase were incorporated. After ensuring the appropriateness of the data collection procedures, the data were analyzed and interpreted. The findings were used to enhance the quality of rater performance and of the proposed systematic rater program. In the final stage, meaningful findings were documented in the form of a short evaluation report and communicated with an audience.

2) Findings of Post-rating

To estimate the effectiveness of the new training program, different measures were conducted, including a post-workshop survey and a post-rating discussion session with six raters. The findings of these measures follow.

(1) Analysis of G-study

Based on the statistical analysis, a G-study was carried out to see the relative variance for each facet. Table 2 shows the results of the G-study (P x R design) for the
holistic scores of the six raters. The results indicate that the variance component for person effect was 0.2215 with 0.0394 standard error, which accounted for 45.31% of the total variance. The variance component for rater effect was 0.0048 with 0.0041 standard error, which corresponds to 0.99% of the total variance. A total of 53.71% of variance came from the interaction effect between person and raters aspects. These findings indicate that the rater effect was minor and major source of variance was examinees and interaction effects. Further analysis for group comparison did not conducted because the amount of variance from rater aspect is almost close to zero.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Variance</th>
<th>%</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0.2215</td>
<td>45.31</td>
<td>0.0394</td>
</tr>
<tr>
<td>R</td>
<td>0.0048</td>
<td>0.99</td>
<td>0.0041</td>
</tr>
<tr>
<td>PR</td>
<td>0.2626</td>
<td>53.71</td>
<td>0.0175</td>
</tr>
<tr>
<td>Total</td>
<td>0.4889</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

(2) FACETS Analysis

Table 3 presents measures of rater performance in terms of both holistic and analytic scoring methods\(^2\). The six raters who participated in this study ranged from -3.11 to -2.09 in severity for the holistic scoring. For analytic scoring, rater severity ranged from -1 to 0.7 on the logit scale. On average, the raters below 0 on the logit scale were more lenient. For the holistic rating, the toughest rater was Rater 3 (-2.09), and the most lenient was Rater 1 (-3.11). For the analytic scoring, R3 (0.71) was still the toughest, and R2 (-1.00) was the most lenient.

Table 3 shows the fit values for the different raters. We can adopt a more conservative measure recommended by McNamara (1996) for misfit [infit is below 0.75 or above 1.3, or Z score is above +2 or below -2], in the holistic rating, Rater 6 showed a misfit rating pattern, or would be considered inconsistent in their own ratings inconsistent rating. For the analytic method, Raters 3 and 4 would be regarded as inconsistent, and Raters 5 showed overfit rating patterns with boundary values 0.75.

With respect to agreement calculated with Rasch analysis, there was a moderate 52.5% agreement among the raters for the holistic method, and a low agreement of 20.6% for the analytic method. The separation indexes were 1.29 and 13.26, and the Chi-square value was significant, implying that the raters in this study differed from

---

\(^2\) The holistic score and analytic scores were estimated differently because the rating scale system was different. The holistic scale has four levels, but the analytic scale has 12 levels.
one another in their ratings (Linacre, 1989).

### TABLE 3
Comparisons of Analysis of Intra-rater Reliability

<table>
<thead>
<tr>
<th>Rater</th>
<th>Severity (Logit)</th>
<th>SE</th>
<th>Infit</th>
<th>Z</th>
<th>Exact Agree%</th>
<th>Severity (Logit)</th>
<th>SE</th>
<th>Infit</th>
<th>Z</th>
<th>Exact Agree%</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>-3.11</td>
<td>0.23</td>
<td>0.78</td>
<td>-1.7</td>
<td>54.7</td>
<td>-0.57</td>
<td>0.04</td>
<td>0.96</td>
<td>-0.5</td>
<td>22.5</td>
</tr>
<tr>
<td>R2</td>
<td>-2.70</td>
<td>0.23</td>
<td>1.05</td>
<td>0.3</td>
<td>52.4</td>
<td>-1.00</td>
<td>0.04</td>
<td>0.95</td>
<td>-0.7</td>
<td>18.8</td>
</tr>
<tr>
<td>R3</td>
<td>-2.09</td>
<td>0.23</td>
<td>1.14</td>
<td>0.9</td>
<td>48.9</td>
<td>0.71</td>
<td>0.04</td>
<td>1.26</td>
<td>3.5</td>
<td>16.1</td>
</tr>
<tr>
<td>R4</td>
<td>-2.55</td>
<td>0.23</td>
<td>0.82</td>
<td>-1.3</td>
<td>56.2</td>
<td>-0.43</td>
<td>0.04</td>
<td>1.23</td>
<td>3.2</td>
<td>23.8</td>
</tr>
<tr>
<td>R5</td>
<td>-3.16</td>
<td>0.23</td>
<td>0.85</td>
<td>-1.1</td>
<td>55.1</td>
<td>-0.18</td>
<td>0.04</td>
<td>0.75</td>
<td>-4.0</td>
<td>22.8</td>
</tr>
<tr>
<td>R6</td>
<td>-2.50</td>
<td>0.23</td>
<td>1.35</td>
<td>2.3</td>
<td>47.8</td>
<td>0.43</td>
<td>0.04</td>
<td>0.82</td>
<td>-2.7</td>
<td>19.6</td>
</tr>
<tr>
<td>M</td>
<td>-2.69</td>
<td>0.23</td>
<td>1.00</td>
<td>-0.1</td>
<td>52.5</td>
<td>0.58</td>
<td>0.00</td>
<td>0.19</td>
<td>2.8</td>
<td>20.6</td>
</tr>
<tr>
<td>SD</td>
<td>0.37</td>
<td>0.00</td>
<td>0.20</td>
<td>1.4</td>
<td>52.5</td>
<td>0.58</td>
<td>0.00</td>
<td>0.19</td>
<td>2.8</td>
<td>20.6</td>
</tr>
</tbody>
</table>

*Note. Holistic: Separation: 1.29, reliability .62, Chi:15.9, df=5, sig=.01
Analytic: Separation: 13.26 (14.53), reliability:.99(1.00) chi:1056.7, df=5, sig=.00*

(3) Interaction (bias) Analysis

Analyses of the interaction between raters and test topics and between raters and assessment criteria were conducted using FACETS. Bias can be detected when raters’ severity level was gradually more lenient or harsh. No interaction effect between raters and test topic was found for holistic analysis. However, for analytic scoring, three statistically significant cases were found (see Table 4). The statistically meaningful bias sizes were 0.17, 0.18, and 0.19 logit. For instance, R4 seemed to score the topic of animal testing more severely than R4’s general scoring pattern. R1 and R2 seemed to rate the topic of globalization more harshly by 0.18 logit or 0.19 logit.

With respect to the interaction of raters and assessment criteria for the analytic scoring method, seven interaction effects were detected. R2 showed a bias on the criterion of focus (0.30 logit). R4 showed a bias on the criteria of support/elaboration (-0.38 logit) and conventions (0.81 logit). R4 severely scored 1.19 logit (absolute value) with conventions than with support/elaboration. R5 also showed interaction with the criteria of conventions (-0.20) and integration (0.20). R6 showed bias for conventions (-0.62) and support/elaboration (0.32).

These findings suggest that all of the control group raters, except R2, showed interaction with the assessment criteria. This is useful information for future training programs, because the suggested criteria can be reanalyzed and modified. In addition, raters should be re-trained on the test topics in order to remove any interaction effects (Linacre, 1989).
<table>
<thead>
<tr>
<th>Rater by</th>
<th>Obs-Exp average</th>
<th>Bias size</th>
<th>Model S.E.</th>
<th>Rater</th>
<th>Test topics</th>
<th>Obsvcount</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>topic</td>
<td>0.20</td>
<td>0.17</td>
<td>0.07</td>
<td>R4</td>
<td>animal testing</td>
<td>150</td>
<td>2.25</td>
<td>.027</td>
</tr>
<tr>
<td></td>
<td>0.22</td>
<td>0.18</td>
<td>0.07</td>
<td>R1</td>
<td>globalization</td>
<td>150</td>
<td>2.47</td>
<td>.015</td>
</tr>
<tr>
<td></td>
<td>0.23</td>
<td>0.19</td>
<td>0.07</td>
<td>R2</td>
<td>globalization</td>
<td>150</td>
<td>2.56</td>
<td>.011</td>
</tr>
<tr>
<td>assessment criteria</td>
<td>-0.69</td>
<td>-0.62</td>
<td>0.10</td>
<td>R6</td>
<td>convention</td>
<td>90</td>
<td>-6.21</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>-0.43</td>
<td>-0.38</td>
<td>0.10</td>
<td>R4</td>
<td>support</td>
<td>90</td>
<td>-3.83</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>-0.22</td>
<td>-0.20</td>
<td>0.10</td>
<td>R5</td>
<td>integration</td>
<td>90</td>
<td>-2.00</td>
<td>.048</td>
</tr>
<tr>
<td></td>
<td>0.23</td>
<td>0.20</td>
<td>0.10</td>
<td>R2</td>
<td>focus</td>
<td>90</td>
<td>3.04</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td>0.36</td>
<td>0.32</td>
<td>0.10</td>
<td>R6</td>
<td>support</td>
<td>90</td>
<td>3.24</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td>0.96</td>
<td>0.81</td>
<td>0.10</td>
<td>R4</td>
<td>convention</td>
<td>90</td>
<td>8.39</td>
<td>.000</td>
</tr>
</tbody>
</table>

(4) Analysis of Rating Accuracy

Rating accuracy was analyzed and estimated by directly comparing the raw scores of the originals with the raters' scores. Table 5 shows the level of exact agreement between the holistic scores of the six individual raters and the original holistic scores. Table 5 shows some interesting results. Both the workshop group and the control group showed more accuracy in the holistic scoring than the analytic scoring, and in group comparisons, the workshop group scored more accurately.

<table>
<thead>
<tr>
<th>Group</th>
<th>Holistic Frequency</th>
<th>Holistic Percent</th>
<th>Analytic Frequency</th>
<th>Analytic Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workshop</td>
<td>134/270</td>
<td>49.62</td>
<td>113/270</td>
<td>41.85</td>
</tr>
<tr>
<td>Control</td>
<td>129/270</td>
<td>47.78</td>
<td>98/270</td>
<td>36.29</td>
</tr>
</tbody>
</table>

V. DISCUSSION

1. Conclusion

The purpose of this study is to develop a new rater workshop program for enhancing rater reliability by evaluating the previous rater training program. Program evaluation theory was adopted in order to conduct this study. Training theory was employed to design the new workshop program. Finally, training effectiveness was estimated with quantitative and qualitative approaches. The summary of the results is following;
i) The EPT training was innovatively reorganized and upgraded. The trainer integrated and updated the training materials to fit the current rating context. Prototype essays were re-selected in consideration of the focus of activity for each phase of the proposed training program. In addition, the training program was designed so that raters could perform evidence-based judgments and solve their rating problems. The workshop lecture provided substantial information, and raters learned how to accurately observe the essay for the appropriate information to match evidence from the rating scale descriptors, and finally how to make a final decision.

ii) It was significant that the summary of the findings of the workshop activity showed evidence that raters successfully understood the training focus and followed the directions. Moreover, a contextualized feedback system was employed via group discussion. Feedback from peers and the trainer was instantly given; however, the consensus process was slightly different depending on the peers. Some groups conducted the consensus process successfully, but some did not. It was suggested that the lecture session provide information about the evidence-based decision process.

iii) Based on the results of the post-rating session, training effectiveness was explored with a quantitative approach by identifying rater reliability. Based on the quantitative findings of the post-rating session, workshop group raters showed a slightly better rating performance in terms of intra-rater reliability, rating accuracy and agreement. Using GENOVA, the variance of rater aspect was estimated, and the findings indicate that the rater effect was minor regardless of scoring method. It is interesting to note that the large amount of interaction among facets accounted for the majority of the errors.

The results of this study provide some theoretical and practical implications to research of a teacher-rater training. First, this study is expected to yield valuable insights for designing a rater training program. The content of the lecture, the scoring methods, and the group activities should be updated and effectively implemented in the future. Secondly, in terms of training method, it seemed that raters who participated in this study were highly motivated by receiving contextualized feedback and encouragement from the trainer. Also, individualized feedback contributes to enhancement of raters’ professional development with high self-awareness. Finally, the findings of this study are important to understanding the rater training program and resolving its practical concerns.
2. Limitations

There are several limitations of this study. First, when research related to rater training effectiveness is conducted, a better research design is required to estimate training effectiveness. Originally, both pre- and post-rating sessions (experimental design) were proposed, but due to the difficulty of recruiting participants, the research design was modified to include only the post-rating session. However, it would have been helpful to compare the quality of rating before and after the workshop, and the evidence would contribute to estimating the effectiveness of the training. Thus, in this study, training effectiveness and data interpretation were somewhat unclear because rater improvement was not clearly identified. For future study, it is suggested that better measures be adopted and implemented.

Second, sampling was one of the limitations of this study. Different participants from the same population were recruited at different research phases. Particularly, for the post-rating session, three of the six raters were concurrently serving as ESL teachers, and they had been involved in additional rating sessions for classroom diagnostic testing. It is likely that practice effects from this additional rating may have reduced training effectiveness. In addition, a small number of participants make it difficult to generalize the results across contexts in some ways.

REFERENCES


University of Chicago.
University of Illinois, Urbana–champaign.

Applicable levels: teacher education
Key words: teacher-rater training model, rater reliability, program evaluation

So Young Jang
Dept. of English Language and Literature
Chung-Ang University
221 Heuk Seok-Dong, Dong Jak-Gu
Seoul 156-756, Korea

Email: clarasoyoung@hanmail.net

Received in June, 2011
Reviewed in July, 2011
Revised version received in August, 2011