PASAA
Volume 48
July - December 2014

Teachers' Reactions Towards Performance-based Language Assessment

Bordin Chinda
Chiang Mai University

Abstract

This research aims at examining the reactions of tertiary EFL teachers towards the use of performance-based language assessment. The study employed a mixed-method research methodology. For the quantitative method, 36 teachers responded to a questionnaire survey. In addition, four teachers participated in the in-depth interviews which were conducted twice, at the beginning and at the end of the semester. The data from the survey revealed that the teachers were aware that the assessments (performance-based and traditional methods) could have both positive and negative impacts on teaching and learning. In addition, the participants in the interviews pointed out that they had rather negative attitudes toward the performance-based assessment used, after it had been implemented for over six years, because there were weaknesses in the assessment especially concerning tasks, rating scales, and rater training. They recommended that the rating scales should be revised and the rater training should be vigorously implemented to ensure the quality of the assessment process.

Keywords: language testing, performance-based language assessment, assessment literacy

Introduction

The English Department at a large public University in the North of Thailand has been responsible for offering English courses to English major students as well as foundation English courses for more than 25,000 students each semester. With the notion of communicative language assessment, the department has incorporated performance-based language assessment (including written and oral assessment tasks) as part of all the foundation English courses, though the core of the assessment has remained in the form of traditional or classical testing. According to Chinda's (2009) longitudinal study, the teachers at the department had different reactions towards the implementation of such performance-based assessment. As a follow-up research study, this present study aimed to investigate the teachers' reactions towards the use of performance-based language assessment after it had been implemented for more than six years prior to the present study since the findings from Chinda's (2009) study indicate that the teachers' attitudes and beliefs influence the way they conduct the assessments in the classrooms. To make sure that the assessment is reliable and valid; therefore, it is crucial to investigate the teachers' attitudes towards and beliefs in the assessments implemented, especially when the assessments have been employed for over six years. It should be noted, however, that this study is not a comparative study. Thus, this paper reports the findings from the data collected in this study, inclusively. Since this study aimed at examining the reactions of teachers towards performance-based language assessment, the following review mainly covers performance-based assessment and its related issues including the differences between performance-based language assessment and traditional language testing, rating scales, and rater training.

Performance - based Language Assessment VS. Traditional Language Testing

Though discrete point items of traditional testing have been the dominate mode of assessment in Thailand (e.g. Ordinary National Educational Test), with the arrival of communicative language teaching, language testing and assessment in many institutions has shifted to focus more on the actual performance of the students (Chinda, 2009; McDonough & Chaikitmongkol, 2007; Watson Todd, 2006); that is, the implementation of alternative assessment. It should be noted that the term "alternative assessment" has been defined differently by different scholars, and different terms have been used to refer to the same concepts. Other terms include authentic assessment, performance-based assessment, continuous assessment and on-going assessment, to name a few. Thus, this review uses these terms interchangeably. Traditional testing emphasises "the rank ordering of students, privileged quantifiable data for isolated, individual test performances, and in general promotes the idea of neutral, scientific measurement as the goal of educational evaluation"; whereas, the "alternative assessment" is based on "an investigation of developmental sequences in student learning, a sampling of genuine performances that reveal the underlying thinking processes, and the provision of an opportunity for further learning" (Lynch 2001, pp. 228 - 229). In addition, Lynch reports that in traditional testing, the testing and teaching are separated activities conducted by separate groups of people of which the students have no access to the criteria and a single score is usually reported. On the other hand, in the alternative assessment, assessment and teaching are integrated with active participation of the students as part of the process of developing assessment criteria and standards. In other words, they are two different cultures. Table 1 below summarises the contrastive features between the classroom-based assessment and classical testing cultures.

Table 1: The two ends of the assessment cultures continuum (Hamp-Lyons, 2007, p. 494)

Classroom-based assessment	**Classical testing**
Fluency-focused	Accuracy-focused
Individual-focused	Group- or 'norm'-focused
Achievement/progress focused	Proficiency-focused
Process-focused	Product-focused
Teachers'/student's voices	Rule-makers' voices
Leads to assessment of learning	Leads to 'teaching to the test'

For performance-based language assessment, in addition, McNamara (1996) states that a defining characteristic of performance testing is that "the assessment of the actual performances of relevant tasks is required of candidates, rather than the more abstract demonstration of knowledge, often by means of paper-and-pencil tests" (p. 6). Moreover, Davies, Brown, Elder, Hill, Lumley, and McNamara (1999) point out that in the performance-based assessment the ability in performing the given tasks completed by a candidate is assessed (p. 144). Tasks, in the assessment of second language performance, are designed to measure learners' productive language skills through performances, which allow learners to exhibit the kinds of language skills that may be required in a real world context (Wigglesworth, 2008, p. 111). Furthermore, Wigglesworth (2008), drawing from McNamara (1996) and Norris, Brown, Hudson and Yoshioka (1998, p. 113) reports that there are three factors which distinguish performance tests from traditional tests of second language: (1) there is a performance by the candidate; (2) the performance is judged using an agreed upon set of criteria; and (3) there is a degree of authenticity with regard to the assessment tasks.

Although, Norris, Brown, Hudson, and Yoshioka (1998) warn that the reliability and validity of alternative assessments must be ensured, Lynch (2003) maintains that since traditional testing and alternative assessments are two different paradigms, they require different reliability and validity frameworks. He asserts that within the alternative assessment approach, "reliability is not necessarily a precondition for validity" as opposed to the traditional testing. Hamp-Lyons (1997) provides a model (Table 2), illustrating the differences between the characteristics of performance/alternative assessments and standardized tests.

Table 2: Characteristics of performance assessments and standardized tests (Hamp-Lyons, 1997, p. 300)

Performance assessment	**Standardized test**
Criterion referenced	Norm referenced
Contextual objectives	Decontextualized objectives
Modifiable	Uniform
Multidimensional	Restricted dimensions
Longitudinal	Pre/post 'snapshots'
Continuous recording	Discontinuous recording
Monitors progress	Static view of achievement
Extensive behaviour sampling	Restricted behaviour sampling
Reflects quality of work	Reflects speed and accuracy
Promotes student learning	Promotes skill in test-taking
Enhances student motivation	Promotes student anxiety
Instructionally relevant	Instructionally independent
Contributes to classroom change	Imposes institutional change
Informs instructional decisions	Justifies bureaucratic decisions
Useful to parents and others	Unhelpful to parents and others

Rating Scales and Rater Training

Another important characteristic of performance-based assessment discussed by McNamara (1996) is that the rater needs to use a rating scale in rating a performance to arrive at a score for that performance (p. 121). In this type of marking, or as it is sometimes referred to as subjective marking, Alderson, Clapham and Wall (1995) stress that the examiners' or raters' job is to assess a task completed by a candidate, for which they need a "rating scale" (p. 107). Therefore, this section explores two major aspects of performance-based language assessment: rating scales and rater training.

A rating scale (or proficiency scale) is a "scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged... The levels or bands are commonly characterised in terms of what the subjects can do with the language... and their mastery of linguistic features" (Davies, et al., 1999, p. 153). Rating scales also represent the most "concrete statement of the construct being measured" (Weigle, 2002). The statements in rating scales are commonly referred to as "descriptors" which describe "the level of performance required of the candidates at each point on a proficiency scale" (Davies, et al., 1999. p. 43). Weigle (2002) identifies three main types of rating scales: primary trait scales, holistic scales, and analytic scales. Nonetheless, the most frequently used scales in an ESL/EFL context are holistic and analytic scales.

With an analytic scale, raters are asked to judge several components of a performance separately, on the basis of traits, criteria, or dimensions of performance. These components are divided so that they can be judged separately rather than expecting the assessor to give a single score for the entire performance (Alderson et al. 1995; Arter & McTighe, 2001; Weigle, 2002). Arter and McTighe (2001) state that analytic scales are used when planning instruction to show relative strengths and weaknesses of a performance, when teaching students the nature of a quality performance, when giving detailed feedback, and when knowing how to precisely describe that quality is more important than speed (p. 25). One main advantage of the analytic scoring method over the holistic counterpart is that it provides a higher reliability (Goulden, 1994). Weigle (2002) also agrees that compared to holistic scoring, analytic scoring is more useful in rater training, and is particularly useful for second-language learners, as it is more reliable. However, she recognises that the rating time that is necessary for analytic scoring takes longer than that of holistic scoring because raters need to make more than one decision for every script. She also adds that a good deal of the information provided by the analytic scale is lost when scores on different scales are combined to make a composite score (p. 120).

In contrast, with a holistic scale, raters are asked to give a judgement on a candidate's performance as a whole, or in other words, a single score for an entire performance based on an overall impression of a candidate's work (Alderson et al. 1995; Arter & McTighe, 2001; Weigle, 2002). Thus, the scale used in this method is sometimes called an impression scale. Arter and McTighe (2001) state that holistic scales are used when the speed of scoring is more important than knowing precisely how to describe quality, when the performances are simple, and when a quick snapshot of overall achievement is the objective (p. 25). This type of scoring method, nevertheless, has been heavily criticised, especially in an EFL/ESL writing assessment context. Furthermore, in the research report for the Educational Testing Service (ETS), Hamp-Lyons and Kroll (1997, p. 28) point out the inherent nature of holistic scoring as being a form of impression marking in a speed dependent manner. They state that "many raters make judgments by responding to the surface of the text and may not reward the strength of ideas and experiences the writer discusses”. This argument is also supported by Shi’s (2001) empirical study, which illustrates that in writing assessment a holistic scoring approach is not an effective method in distinguishing the salient differences of students’ performances. From the rater’s comments, Shi observes that holistic rating raises questions about the construct validity because the rater’s comments demonstrate that they had different understandings of what constitutes good writing. In a more recent study, Barkaoui (2007) had four EFL writing teachers rate 32 essays, without any formal training. These essays were written by intermediate EFL university students in Tunisia under exam-like conditions, of which four were used for the think-aloud sessions. Both analytic and holistic scales were used. Interestingly, contrary to the concept that a holistic scale yields a lower level of reliability than a multiple trait scale, Barkaoui found that when the essays were rated holistically, a higher level of score reliability was achieved. He also reported that analytic scoring resulted in high rater variability and that more ratings were required in order to achieve acceptable dependability indices. In a different context, Iwashita and Grove (2003) studied the assessment of the speaking component of the Occupational English Test (OET) for health professionals in Australia. Iwashita and Grove examined the relationship between analytic and holistic scales used in this testing system where a combined analytic-holistic assessment scale was used. Their study included 13,488 assessments (consisting of assessments by 29 raters) which were collected over eight years. The data was analysed by means of the many-faceted Rasch model programme, FACETS. The results from the analysis of the rating patterns using both analytic and holistic scales suggested that the overall scores did not accurately reflect the candidate’s ability, and the analytic rating could be overrated. Iwashita and Grove concluded that it was possible that using a single holistic criterion may be more accurate and efficient than the combined scale.

In terms of a rater training, Alderson et al. (1995) point out that one of the most important issues to consider in teacher assessment is rater monitoring. They also state that training the examiners or raters could provide them with "competence and confidence" (p. 128). A rater training prepares raters for the task of judging the candidate's performance. It mainly involves the process of familiarising the raters with the test format, test tasks, rating scales, and exemplar performances at each criterion level (Davies, et al., 1999, p. 161). In order to improve the quality of rater-mediated assessment, McNamara (2000) emphasises the moderating meeting scheme, providing initial and on-going training to raters. For the process of conducting rater training, McNamara proposes that the rater training must involve individual raters independently marking a series of different levels of performance. Then, in groups, they have to share their marks with the other raters. The differences are noted and discussed in detail by referring to the interpretation of the different levels of the descriptors of the individual raters. The purpose of the meeting is to try to bring about a general agreement on the relevant descriptors and the rating categories. Similarly, Alderson et al. (1995) suggest that the first stage of the meeting should be devoted to discussing the consensus scripts to find out if all raters agree on the marks that have been given, and to work out why problems exist, if they do not agree. The aim of this activity is to help all raters match the marks of the original committee. Thus, the committee's consensus scores should not be indicated on the scripts. They explain that the raters should not be shown the decisions made by the committee "to prevent examiners from being influenced by the original committee's reasoning before they have had a chance to try out the scale and think for themselves" (p. 112). The consensus scripts are those scripts that represent "adequate" and "inadequate" performances, as well as scripts that present some common problems which raters often face, but are rarely described in rating scales. The raters should try out the rating scale on the consensus scripts, which are given before the meeting. After that, the problematic scripts should be presented, together with guidelines on what raters should do in these cases. Then, further practice in marking should be provided with another set of scripts. However, it should be noted that rater training may not ensure the inter-rater reliability but rather the intra-rater reliability. In her study, Weigle (1994) investigated the effects of training on the raters of ESL compositions using both quantitative and qualitative methods. In this study, Weigle included 16 raters, of which half were inexperienced raters (who were the focus of the study). The data was collected before, during and after the training sessions. The data revealed that the training helped the inexperienced raters to understand and apply the rating criteria. The training also brought these raters "more or less in line with the rest of the raters" (p. 214). However, a new insight was revealed when Weigle later applied the multifaceted Rasch measurement to analyse the data. From the analysis, Weigle (1998) found that "rater training cannot make raters into duplicates of each other, but it can make raters more self-consistent" (p. 281).

Lumley and McNamara (1995) also report that the results of rater training are not long-lasting. Lumley and McNamara compared the test scores from the Occupational English Test administered in Australia which were obtained from two rater training sessions, 18 months apart, and a subsequent operational administering of the test (about two months after the second training session). They employed the multifaceted Rasch measurement and found inconsistencies and changes in the raters' behaviour between the rater training sessions and the actual test administration, especially from the second training session and the operational administration. Lumley and McNamara, thus, suggested that rater training should be conducted at every administration of the test.

In conclusion, alternative assessment has become an umbrella term used to refer to performance-based assessment as well as the "alternatives" to traditional discrete-point tests (Fox, 2008). Drawing from the above discussions, great care is needed when implementing performance-based language assessment, especially for the purposes of ensuring its quality and impact on learning. The quality of such assessment could be initially controlled by the verification of the reliability of the rating scales and the on-going rater training. It should be noted, however, that in making a decision on which paradigm to adopt, those involved in making the decision, in which classroom teachers must be included, should initially take the purposes of teaching and learning into consideration. Arguably, when the purposes of teaching and learning focus on the construction and administration of standardized or traditional tests in which teaching and testing are separated, the traditional test method is likely to be chosen. Unfortunately, in this circumstance, the potentials of performance-based assessment are neglected. On the other hand, when performance-based assessment is being adopted, teachers are not well prepared to employ it. In this circumstance, the implementation of the performance-based assessment could cause a number of problems among the teachers. Therefore, this study aimed to examine the teachers' reactions, including their attitudes, beliefs, and practices, toward performance-based assessment after it had been implemented for a number of years.

Research Methodology

In terms of the research methodology, a questionnaire survey, with a 5-point Likert Scale and multiple-choice items, was administered to the teachers teaching the foundation courses at the department. The questionnaires used in the present study was from Chinda's study (2009) (see the appendix for the questionnaire) since this study had the same purposes including the determination of the attitudes toward, beliefs in, and practices in the performance-based and traditional assessments. Moreover, both studies were conducted at the English Department at the same university where the assessments used (in the former study) had not undergone major changes. The questionnaire asked the respondents to indicate how much they agree or disagree with the statements concerning their views regarding and experience with assessment, personal assessment experience, and opinions of assessments in general. The data from the questionnaire was analysed using IBM SPSS Statistics Version 20. Because the questionnaire was adopted from Chinda's study (2009), which was already piloted, the questionnaire was not piloted in the present study. However, the reliability coefficient was computed for this study. The reliability coefficient of the survey was .761 (cronbach alpha), which is considered a reliable score. For the qualitative part, in-depth interviews were conducted with six teachers. The interviews were conducted at the beginning of the semester and at the end of the semester. However, two teachers decided not to participate in the second interview. Therefore, the data from four teachers (Pranee, Rattana, Nisa, and Supee) are reported in this paper. These four teachers have been teaching foundation English courses for a different number of years, in addition to teaching various other English major courses. The names of the participants used in this paper have been changed for confidential purposes.

Findings and Discussion

This section is divided into two parts: overviews of the data accumulated from both the quantitative method and the qualitative method. The quantitative data described the attitudes of the teachers in terms of their views toward language assessment in general, views toward the assessment used at the department, and their personal assessment experience. For the qualitative part, the discussions focus on the participants' reactions towards performance-based language assessment, which has been employed at the department.

1 Teachers' attitudes toward assessment and assessment practices: results from questionnaire surveys

This section illustrates and discusses the results from the questionnaire survey. Table 3 reveals the demographical information of the questionnaire respondents. From the table, there were 56 teachers teaching the foundation English course under investigation. However, according to the table, only 36 questionnaires were returned (63.23%). With the limited space of this paper, with regard to the Likert Scale items, only the ones with very high agreement (scores between 5.00 and 4.21) and high agreement (3.41 and 4.20) have been reported. The table below illustrates the demographic information with regard to the questionnaire respondents. According to Table 3, the majority of the respondents were female aged between 31 – 40 with an MA (English) and MA (Education) educational background. Most of them had more than 10 years of teaching experience. In addition, almost about the same number of teachers had and did not possess training in testing in their pre-service training.

Table 3: Demographical information

		Frequency	**Percent**
Gender	Female	31	86.11
	Male	5	13.89
Age group (year)	20 - 30	3	8.33
	31 - 40	13	36.11
	41 - 50	5	13.89
	Above 50	15	41.67

		Frequency	Percent
Qualification	MA (English)	14	38.89
	MA (Education)	16	44.44
	MA (Linguistics)	3	8.33
	PhD (Testing)	1	2.78
	PhD (Literature)	1	2.78
	PhD (Education)	1	2.78
Work Experience (year)	1-3	2	5.56
	4-6	5	13.89
	7-10	4	11.11
	More than 10	25	69.44
Training in Testing	No	19	52.78
	Yes	16	44.44

1.1 Views toward language assessment in general

Since assessing students has become a very important part of teachers' work, this section aims to understand teachers' viewpoints on assessment in general. From the attitudes of the respondents, it can be concluded that these teachers had rather positive attitudes toward assessment in general. More importantly, they are aware of the fact that assessment has an impact on learners. Drawing from Alderson and Wall (1993), the impact of assessment could be positive or negative. According to the data (Table 5), the respondents thought that assessment could have a positive impact on students as they highly agreed that the assessment results have an important effect on the students' self-concept; consequently, students try to achieve their best. Though they were aware that assessment could have negative effects on students, they believed that these effects could create positive effects, as the data indicated that the respondents agreed that assessment creates competition and they felt that assessment motivates learning.

Apart from realizing the impact of assessment on learners, the respondents were aware that assessment could have an impact on teaching. According to the data, the teachers highly agreed that assessment results are important for instruction. They also agreed that assessment highlights each student's strengths and weaknesses. In other words, the teachers agreed that the results from the assessment could be used to guide their teaching, as the results could provide them with diagnostic information about the ability of their students.

Table 4: Teachers' views toward assessment in general

	Mean	**SD**
Assessment creates competition	3.94	.715
Assessment motivates learning	3.92	.806
Assessment results affect student self-concept	4.31	.577
Assessment highlights each student's strengths and weakness	4.19	.749
Assessment improves learning	3.97	.878
Assessment results are important for instruction	4.39	.645

1.2 Views towards language assessment at the department

As the assessments used at the department consist of both performance-based and standardized classical tests, this part of the questionnaire aimed to understand the respondents' reactions to both types of assessment with the focus on their personal reactions toward the assessment tasks. The data discussed in this section was drawn from part of the questionnaire employing multiple-choice items.

Table 5: Teachers' views toward the assessments used at the department

		Frequency	**Percent**
Strengths	Diagnose students	15	41.7
	How well I taught	3	8.3
	Focus on my teaching	1	2.8
	Represents students' abilities	15	41.7
	Others	2	5.6
Weaknesses	Too long to score	4	11.1
	Unreliable	10	27.8
	No detailed feedback	18	50.0
	Allowing cheating	3	8.3
	Others	1	2.8

From the survey, it appears that the respondents agreed that the main strength of the assessments, including both standardized classical tests and performance-based assessments, used at the department was they can represent students' abilities. These results could indicate that these teachers realized the benefits of performance-based assessments, though it was not the major assessment implemented at the department, used in a classroom context. This is of relevance because Wigglesworth (2008) has pointed out that with this kind of assessment, students are allowed to perform the specific language skills which may be required in real world activities. In addition, the teachers in this study have found that through this kind of assessment, they can diagnose students' language ability. In other words, the teachers might believe that having included performance-based assessment in this assessment context could have provided this benefit. In contrast to their views towards assessment in general (as discussed above), the respondents did not think that the assessments used at the department could provide them with beneficial information for instructional purposes. The reason for this could be related to the fact that the majority of the assessment tasks that were used involved standardized classical testing. As Hamp-Lyons (1997) has pointed out that the standardized test is “instructional independent” (p. 300), the teachers, therefore, believed that the assessments used at the department did not offer them information on the topic of “How well I taught”. Also they did not think that the assessments represented the students’ abilities because the results from the standardized tests could not provide information on the “multi-dimensional” abilities of the students (Hamp-Lyons, 1997, p. 300) (see also Table 2).

Nevertheless, the questionnaire respondents pointed out that the main weakness of the assessments used in the department was that they do not allow for teachers to give detailed feedback to the students. This is because of the nature of the assessments used, where there is a large number of students in one class that must be accommodated. There are approximately 36 students in one class. Therefore, it would require that teachers give a great deal of time to provide detailed feedback for each student assessment task, especially for written tasks Furthermore, the respondents agreed that the assessments allow students to try their best. The main reason the teachers agreed with this aspect could be due to the fact that performance-based assessments have been adopted. Approximately six years prior to the present study, only traditional multiple-choice midterm and final exams were implemented at the department (see Chinda, 2009 for more details) for the foundation English courses. When performance-based assessments, including oral and written tasks, were introduced, teachers might have realized that within this assessment context, students are not under “exam conditions”. Therefore, they do not have that pressure and they tend to perform their best.

1.3 Personal assessment experience

The main objective of this part of the questionnaire was to investigate the assessment practices of the teachers.

Table 6: Teachers' assessment experience at the department

	Mean	SD
Consistent when scoring	4.00	.840
Discuss scoring issues with other teachers	3.74	1.067
Need verities of assessment	4.08	.806
Have enough assessment knowledge	3.44	1.081
Discuss writing assessment issues with other teachers	3.81	1.064
Discuss oral assessment issues with other teachers	3.67	.986
Discuss exam issues with other teachers	4.03	.696
Scoring writing assessments takes a lot of time	4.44	.773
Marking exams takes a lot of time	3.94	1.120

As shown in Table 6, the respondents highly agreed that scoring writing assessments takes a lot of time. Since the rating scales used in the foundation courses at the department were analytic scales, this finding is in line with Weigle's (2002) warning about the analytic scales in which it was stated that analytic scoring takes longer than that of holistic scoring. With a lower score from the survey, the respondents agreed that marking exams also took a lot of time. This could be due to the fact that each teacher had to mark exams for more than 35 students for each section they taught and the majority of teachers taught about five sections. Moreover, with a similar level of scores, the teachers agreed that they had discussed scoring, writing assessments, oral assessments, and exam issues with their colleagues. Also they agreed that they needed verities of assessments to assess their students. This result illustrates that these teachers agreed with the use of alternative assessments as this type of assessment requires different methods to assess the students (Norris et al, 1998).

It is very interesting, nevertheless, to point out that the respondents believed that they have been consistent in scoring their students performances. In other words, they believed that they had intra-rater reliability. Unfortunately, this research did not include an intra-rater reliability study. It would be interesting to see whether the teachers have as high a level of intra-rater reliability as they reported. Nevertheless, though intra-rater reliability was not included in the present study, Chinda's study (2009) examined the inter-rater reliability of the teachers in the department. According to the study, which employed Rasch measurement analysis, teachers in the department were quite different in their degree of severity when they rated students' performances (p. 124).

Furthermore, the teachers also thought that they had an adequate understanding of how to construct assessments. It should be noted that the assessments used in the department are constructed by a group of committee members who are the coordinators of each course. Some other teachers might be requested to help with writing a certain amount of items and some teachers might be requested to give feedback. Nevertheless, according to the personal information in Part I of the questionnaire, only about half of the respondents had training in language testing and assessment in their pre-service training.

2 Investigating the reactions of teachers towards the use of performance-based assessment: findings from the interviews

This section reports on the attitudes, beliefs, and the practices of four teachers regarding the assessment practices being used in the department. The teachers had some experience applying the assessment method prior to this study but with a different number of years of teaching experience with regard to the courses. Moreover, the data is drawn from two interviews. The first one was conducted at the beginning of the semester (that is before they assessed the students) and the second one was conducted near the end of the semester (that is after they had assessed the students).

2.1 Reactions towards performance assessment

From the interviews with the participants, it can be stated that they had different attitudes and beliefs towards the use of performance-based assessment (including written and oral assessment). First of all, the first participant believed that performance-based assessment was beneficial to the student. Nisa believed that written and oral assessments during the semester helped reduce students' levels of stress with regard to the final exams. In Interview 1, she stated that in the foundation course she was teaching, there were quizzes and assignments that allowed her to consecutively assess students throughout the semester. In this way, she believed that students would not feel too stressed out during the final exam. Nevertheless, Nisa argued that for this course the rater reliability could not be controlled, as there was a lack of rater training. In addition, teachers might not give feedback to students. Furthermore, Nisa pointed out that the overall grading system being used by the department might not accurately reflect students' performance or ability.

However, due to a large class size, other participants, who preferred standardized classical tests, did not think that performance-based assessment could be implemented reliably. In the first interview, Pranee stressed that because of a large class size, it was not possible to do classroom-based assessment. Thus, she preferred traditional multiple-choice exams because they were less subjective and more reliable than their performance counterpart.

That is, Pranee was concerned with the practicality of the assessment. Thus, Pranee stressed that standardized tests would be the answer to this issue. In other words, she thought that standardized classical tests were more practical. However, she pointed out that there must be a group of experts brought in to monitor the quality of the exams. Because the performance-based assessment has caused problems with the grading, Pranee stated that there have been efforts to make the assessment less subjective by adding more multiple-choice exams into the courses. That means that the scores for performance assessment have decreased. She believed that this is fair for students because everyone would be using the same exam and teachers did not have power over the scores. Nevertheless, Pranee pointed out that the exam did not cover every unit, especially the reading sections. The reading passages in the exam could only cover limited themes, which might cause bias as some students might not feel comfortable with a certain topic. She argued that though the exams are biased, the exam writers tried to cover all grammar concepts. However, because of the limited number of exam items, it was not possible to cover everything. Pranee also pointed out that the requirements for the written assessment tasks were too mechanic, which did not provide students enough freedom to use their creativity.

Similarly, Ratana agreed that oral assessments allowed for a great deal of subjectivity. In other words, the scores were not reliable as a result of the low inter-rater reliability. In Interview 1, Ratana said that she found that there was a great deal of subjectivity in rating students' oral assessments. She stressed that "We can't be sure if teachers follow the rating scales strictly, and whether the grades truly reflect the students' ability." However, Ratana pointed out that for the speaking courses, teachers had a reduced workload in terms of marking. For a writing course, Ratana expressed that it might be too difficult for teachers as there were too many students in one class. She said “Teachers have to grade many papers which increases the teachers’ workload”. She suggested that there should be fewer assignments which would allow teachers to give more detailed feedback.

2.2 Reactions towards rating scales and rater training

In terms of the analytic rating scales implemented at the department, the participants had different attitudes toward their clarity and how teachers interpreted the criteria. Some participants thought that the criteria in the rating scales were clear; whereas some thought they were too broad. Nisa pointed out in Interview 1 that the rating scales were clear enough for her to use when rating students’ performances. In the second interview, Nisa confirmed that the criteria used in foundation English courses were clearer when compared to the ones in the past, but the problem remained that teachers did not follow the criteria strictly. However, she highlighted that students were not ready for the assessments, especially for the courses requiring written assessments. Moreover, Nisa identified that teachers might not have a similar understanding of the criteria. She said, “for the criterion ‘relevant’, teachers have to interpret this term … We can’t be sure if all teachers interpret the term in the same way when rating students’ performance”. Nisa, furthermore, stressed that rater training should be implemented to improve consistency among teachers’ understanding of the rating scales. In other words, trainings or workshops on using the rating scales should be provided. She pointed out that there was a training session once, but it was not effective and sufficient.

Concerning rating, Pranee noted that native-speaker teachers and some Thai teachers might be too lenient with their marks, which she thought was not fair. Similar to Nisa, Pranee stressed that teachers should follow the rating scales to make the assessment as fair as possible. However, she admitted that because of her lack of knowledge in assessment she was not sure whether the present criteria could truly reflect students' levels of ability. She added that the mechanical nature of grading written assessments was too rigid. Although, it helped with reliability, Pranee pointed out that it blocked students' creativity. She believed that students learn best when they enjoy the writing tasks. She emphasised that grades were not important to her, but she wanted students to be inspired to learn; that is, she wanted students to acquire life-long learning benefits. In the second interview, Pranee stated that the rating scales for the written assessment were clear, and that the exams were easy to mark. However, the rating scales for the oral assessment were too general and students might not understand them. Thus, students might not understand when they received a lower grade than they had expected to be given.

Unlike Nisa and Pranee, Supee stated that the criteria for the oral assessments were difficult to follow while the criteria for the written assessments were clear. Supee stated in the first interview that there have been attempts to make assessments more systematic. She stressed that she could now understand the criteria better, especially the rationale behind each criterion. However, she pointed out that although teachers have been provided with rating scales, it was very difficult to follow them, especially for oral assessment since there were many criteria in the rating scales to consider. Supee argued that because the students had different language problems, it was very difficult to use the criteria and rate students appropriately. Supee also agreed that teachers were not familiar with testing theories and that they might misunderstand the criteria. In Interview 2, Supee pointed out there is a lack of experts in language testing and assessment at the department. Therefore, teachers were not familiar with testing theories, and instead were more familiar with "traditions".

When an expert (referring to the researcher) introduced theories and revised the rating scales, Supee found that she was confused. Yet she found that the rating scales (which were analytic scales) decreased the amount of time spent in the rating process. Nevertheless, Supee noted that there were possibilities for teachers to misinterpret the descriptors. Moreover, Supee stressed that experts should be involved in the assessment development process. She said “We have to invite the experts. We can’t achieve the goal if we don’t know the directions. We need the experts. The department should also support staff to study in this field. These people will come and help us.”

Likewise, Rattana was aware of the weakness of the rating scales used in the foundation courses. In Interview 1, she suggested that the criteria should be less subjective and assess more aspects of students’ language ability. In addition, the criteria should be practical and fair when used with a large number of students. Rattana stressed that teachers should rate students’ written assignments carefully with more detailed feedback. She recommended that there should be workshops for teachers on how to give effective feedback, and that there should be individual consultations for students. In Interview 2, Ratana maintained that for the writing courses, the rating scales needed to be improved, as the criteria were unclear and contained a great deal of subjectivity. She said “This isn’t fair for students.” She explained that an A student in one section might be a B student in another section because of rater differences. Therefore, she strongly agreed that the rating scales should be revised.

In summary, since the implementation of the performance-based assessment, despite the fact that the majority of the participants agreed that this type of assessment could provide benefits to students and teachers, they did not feel that this kind of assessment has been successfully implemented at the department. This might be due to the incompatibility that exists with regard to the previous practice. In other words, the participants in the study were more familiar with multiple-choice exams or standardized classical tests, which have been widely used in all levels of education in Thailand (Prapphal, 2008). Moreover, performance-based assessment is more complex than multiple-choice exams as it involves the construction of the rating scales and the vigorous implementation of a rater training process (as discussed in detail in the literature review above). Therefore, though the participants realized the advantages of the performance-based assessment, these two attributes may have resulted in the emergence of this negative attitude toward this type of assessment. It should be noted that the use of analytic rating scales was only implemented about three years prior to the present study, as a result of Chinda's study (2009). Since then, the rating scales used for foundation English courses at the department have been analytic because teachers were aware of their advantages. However, the data revealed that after they have been used for a few years, analytic scales, which were not the same as the previous scales, might be too complex and time consuming for the teachers to employ in the foundation courses. In other words, analytic scales might not be practical in the present context due to the large class size (about 36 students in each class) and the number of teachers teaching the course (approximately 56 teachers for each course). Therefore, to increase the reliability and validity of the performance-based assessment, the participants agreed that more systematic and frequent rater training should be provided at the department.

Conclusions

Though performance-based language assessment has been considered an authentic method in assessing students' abilities, the main-stream traditional method of testing has remained the most influential way of testing in Thailand. With the aim of finding out the reactions of EFL teachers teaching foundation English courses toward the implementation of performance-based assessment where the main assessment method is traditional testing, the present study employed both quantitative and qualitative research methods. The data from 36 teachers, who responded to a questionnaire, revealed that the teachers who participated in the study had both positive and negative feelings toward assessments in general and specifically to the ones used at the department. Also, they believed that assessments could have both positive and negative impact on both teachers and students; that is, on teaching and learning. However, the participants in the interviews had mixed attitudes toward the assessment used at the department, as they were aware of the weaknesses of the assessment. For the performance-based assessment which has been implemented for over six years, the participants stressed that the assessment tasks and the rating scales were not as reliable as they would have expected them to be. Thus they recommended that both the assessment tasks and the rating scales should be revised with the supervision of language testing experts. Furthermore, the participants believed that to make certain the reliability of the scores, rater training should be vigorously implemented.

The Author

Bordin Chinda is a lecturer at the English Department, Faculty of Humanities, Chiang Mai University. His research interests include impact/wash-back studies, performance-based assessment, innovations in language education, professional development, and English for Academic Purposes. Bordin holds a PhD in Language Testing and Assessment from the University of Nottingham, UK.

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115-129.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks, CA: Corwin Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. Assessing Writing, 12(1), 86-107.

Chinda, B. (2009). *Professional development in language testing and assessment: A case study of supporting change in assessment practice in in-service EFL teachers in Thailand*. Unpublished PhD Thesis, The University of Nottingham, UK.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing* (Vol. 7). Cambridge: Cambridge University Press.

Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *The Journal of Research and Development in Education, 27*(1), 73-82.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing, 14*(3), 295-303.

Hamp-Lyons, L. (2007). The impact of testing practices on teaching: Ideologies and alternative. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (Vol. Part I, pp. 487-504). New York: Springer.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 - writing: composition, community, and assessment*. Princeton, NJ: Educational Testing Service.

Iwashita, N., & Grove, E. (2003). A comparison of analytic and holistic scales in the context of a specific-purpose speaking test. Prospect, 18(3), 25-35.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications for training. Language Testing, 12(1), 54-71.

Lynch, B. K. (2001). The ethical potential of alternative language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Language Testing in honour of Alan Davies* (Vol. 11, pp. 228-239). Cambridge: Cambridge University Press.

Lynch, B. K. (2003). *Language assessment and programme evaluation*. Edinburgh: Edinburgh University Press.

McDonough, K., & Chaikitmongkol, W. (2007). Teachers' and learners' reactions to a task-based EFL course in Thailand. *TESOL Quarterly, 41*(1), 107-132.

McNamara, T. (1996). *Measuring second language performance*. London: Addison Wesley Longman Ltd.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu: University of Hawai'i.

Prapphal, K. (2008). Issues and trends in language testing and assessment in Thailand. *Language Testing, 25*(1), 127-143.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. Language Testing, 18(3), 303-325.

Watson Todd, R. (2006). Continuing change after the innovation. *System*, (34), 1-14.

Weigle, S. C. (1994). Effects of training on raters of English as a second language composition: Quantitative and qualitative approaches. Unpublished PhD Thesis, University of California, Los Angles.

Weigle, S. C. (1998). Using FACETS to model rater training effects. Language Testing, 15(2), 263-287.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Wigglesworth, G. (2008). Task and performance based assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7, pp. 111-122). New York: Springer Science+Business Media.

Appendix

Questionnaire: Teachers' Views of Foundation English Assessments

Your personal information

Your gender:

❍ Female ❍ Male

Your age:

❍ 20 – 30 ❍ 31 – 40 ❍ 41 – 50 ❍ above 50

Your academic qualifications (please answer all that apply):

❍ Bachelors Degree in ....

❍ Masters Degree in ....

❍ PhD Degree in ....

Number of years you have been teaching English (including before teaching at Chiang Mai University):

❍ 1 – 3 ❍ 4 – 6 ❍ 7 – 9 ❍ 10 and above 10

Have you had any formal training related to language testing and assessment?

❍ No ❍ Yes, please specify ....

Part 1

For each item, please circle **ONE** of these alternatives which you **most** agree with.

1. The main **strength** of the way we assess our students is:
   a. it indicates which students are doing well and which ones are doing poorly.
   b. it helps me find out how well I have taught my students.
   c. it helps me to focus on my teaching.
   d. it represents the language abilities of the students.
   e. other (please specify) ....

2. The main **weakness** of the way we assess our students is:
   a. it takes too long to score.
   b. the results are unreliable.
   c. it does not allow us to give detailed feedback.
   d. it allows for a great deal of cheating.
   e. other (please specify) ....

3. In relation to the students, the way we assess our students:
   a. motivates them to work harder.
   b. provides them with valuable learning experiences.
   c. provides them with enough feedback which will further their learning.
   d. allows them to try to achieve their best.
   e. other (please specify)

Part 2

Please indicate how far you agree or disagree with each of the following statements. **Circle the number** corresponding to your selection. Please use the following scale:

5 = Strongly Agree (SA) 4 = Agree (A) 3 = Uncertain (U)
2 = Disagree (D) 1 = Strongly Disagree (SD)

Your personal assessment experience

4. I spend more time scoring and recording than giving feedback.	5	4	3	2	1
5. I am consistent in scoring students' written and oral performances.	5	4	3	2	1
6. I often discuss the issue of consistency in scoring students' written and oral performances with other teachers.	5	4	3	2	1
7. I need a variety of assessment methods to assess my students.	5	4	3	2	1
8. I have adequate understanding of how to construct assessments.	5	4	3	2	1
9. I often discuss the issues in assessing the **written assignments** with other teachers.	5	4	3	2	1
10. I often discuss the issues in assessing the **oral projects** with other teachers.	5	4	3	2	1
11. I often discuss the issues concerning the **exams** with other teachers.	5	4	3	2	1
12. Scoring the **written assignments** takes a lot of my time.	5	4	3	2	1
13. Scoring the **oral projects** takes a lot of my time.	5	4	3	2	1
14. Marking the **exams** takes a lot of my time.	5	4	3	2	1

Your views toward assessments in general

15. Assessments create competition among students.	5	4	3	2	1
16. Assessments motivate students to learn.	5	4	3	2	1
17. Assessments take up more time and effort than they are worth in an instructional sense.	5	4	3	2	1
18. Assessment results have an important effect on student self-concept.	5	4	3	2	1
19. Assessments highlight each student's strengths and weakness.	5	4	3	2	1
20. Performance-based assessment (i.e. written & oral tasks/projects) is better than traditional final/midterm exam.	5	4	3	2	1
21. Assessments improve students' learning.	5	4	3	2	1
22. Students dislike being assessed.	5	4	3	2	1
23. Assessment results are important for instruction.	5	4	3	2	1

Your views toward foundation English assessments

24. The oral presentation projects are interesting for students.	5	4	3	2	1
25. The performance results evaluate the instructional units to see if they worked.	5	4	3	2	1
26. There are clear performance criteria for the oral performances.	5	4	3	2	1
27. Students learn something useful from doing the projects.	5	4	3	2	1
28. The projects motivate students to learn.	5	4	3	2	1
29. The students' oral performances represent their ability in speaking.	5	4	3	2	1
30. Most of my students are anxious about the projects.	5	4	3	2	1
31. The projects are interesting for me.	5	4	3	2	1
32. The scoring for the oral performances is balanced and fair.	5	4	3	2	1
33. The exams are interesting for students.	5	4	3	2	1
34. The exam results evaluate the instructional units to see if they worked.	5	4	3	2	1
35. The exams diagnose the strengths and weaknesses of individual students.	5	4	3	2	1
36. There are clear specifications for the exams.	5	4	3	2	1
37. Students learn something useful from doing the exams.	5	4	3	2	1
38. The exams motivate students to learn.	5	4	3	2	1
39. The students' exam performances represent their abilities in reading, grammar and vocabulary.	5	4	3	2	1
40. Most of my students are anxious about the exams.	5	4	3	2	1
41. The exams are interesting for me.	5	4	3	2	1
42. The scoring for the exams is balanced and fair.	5	4	3	2	1

Part 3

54. What improvements do you want to be done to English 101 assessment?

....
....

55. What could be done to improve the assessment situation at the division?

....
....

Thank You Very Much