CHAPTER ONE

INTRODUCTION

Background of the Study

In many EFL settings, reading ability and grammatical knowledge have been considered the two main fields in large scale examinations. However, with the transition to communicative language teaching, helping learners' build up their speaking proficiency and assessing their oral proficiency in an efficient way have become topics of greater emphasis among investigators, educators, and teachers. The reason why evaluations of learner performance have drawn increased attention from people in language education is that score results not only reflect the effectiveness of instruction but are typically used as a threshold for entering a school or applying for a job. Thus, scoring procedures must be both valid and reliable.

Many large scale language proficiency tests, such as TOEFL iBT and GEPT, have already integrated oral proficiency into their exams. Another example, specially designed for testing examinees' discourse and overall oral proficiency, is the PhonePass exam. Test takers of PhonePass have to "read aloud, repeat sentences, say words, and answer questions" via a telephone (Brown, 2003, p. 145). Since the score is calculated by computer, Brown (2003) indicated that the computer-assisted assessment, the PhonePass test, has a higher reliability and correlation statistic than human scoring.

This result also highlights one of the most controversial issues in assessing speaking, that is, lower rater reliability of human scoring. However, the vast majority of speaking assessment still depends on human scoring nowadays. Alderson, Clapham, and Wall (2002) indicated that there are many factors which may interfere with rater reliability and also reduce rater reliability, such as problems with the interpretation of rating scales and time pressures. Fortunately, some researchers have shown that rater training can be used a way to improve rater reliability (Alderson, Clapham & Wall, 2002; Brown, 2003; Lumley & McNamara, 1995; Weigle, 1994, 1998).

Rater reliability has two aspects, inter-rater reliability, and intra-rater reliability. According to Luoma (2005) intra-rater reliability, termed "internal consistency" (Luoma, 2005, p. 179), means that raters are able to give similar scores for the same test over a period of time. Alderson et al. (2002) indicated that raters are still considered reliable if only some of the scores they give are different, but too much variation on different occasions can be questionable. As for inter-rater reliability, Brown (1999) explained that two or more raters have to reach consensus in their judgments about the examinee’s performance. In other words, inter-rater reliability refers to the extent that how raters’ scorings are different from each other. Alderson et al. (2002) considered that these two kinds of reliability complement each other. They indicated that intra-rater reliability can also be monitored when inter-rater reliability is being checked; that is because any agreement among raters will be limited by the internal consistency of each rater. In other words, if a rater can not remain self consistency, his or her variation may also influence the result of inter-rater reliability.

High rater reliability can ensure the fairness of a test. However, attaining inter-rater reliability seems to be considered as a more difficult goal to be achieved. Lunz and Stahl (1990) indicated that raters always have their unique standards of their own, and it is quite difficult to truly alter their principles or have them sacrifice their own voices. Although it is not realistic to expect all raters to always match one another; it is essential that at lest each rater should try to match the standard at all times (Alderson et al., 2002). It should be noted that achieving inter-rater reliability does not mean forcing raters to agree with each other completely, but entails reducing the wild differences in raters’ scores (Luoma, 2005).

Many researchers have suggested that one of the good ways to achieve inter- and intra-rater reliability is through rater training (Alderson, Clapham & Wall, 2002; Brown, 2003; Lumley & McNamara, 1995; Weigle, 1994, 1998). They mentioned that rater training did help raters to maintain internal consistency and consistency across raters. In addition, this kind of training also assists raters to mutually construct and interpret convincing rating criteria or scales. Jacobs et al. (1981) also agreed with the usefulness of rater training. They considered that training not only helps raters to reduce extreme differences in scoring which are outliers in terms of raters' harshness or leniency, but also ensures raters' consistency of applying the criteria. Further, Elder, Iwashita, and McNamara (2002) observed that rater training can modify raters' expectation of task demands and clarify their rating criteria, thereby reducing rater variability. Weigle (1998) even emphasized on the difficulty of deriving usable measures of examinees' ability from any untrained raters.

Rater training refers to the processes in which assessment criteria are introduced to the raters, after which they are required to rate several samples based on the criteria (Elder, Knoch, Barkhuizen, Knoch, & Randow, 2007). Luoma (2005) and Underhill (1991) proposed that rater training should last several days and there are even re-training sessions. Rater training should be carefully designed, especially for large scale and high stakes assessment situations, and ongoing training is also needed (Congdon & McQueen, 2000). However, many researchers still consider regular training sessions to be quite impracticable because they are too time consuming for classroom teachers. Quite simply, teachers do not have sufficient time for regular rater training (Hamilton, Reddel, & Spratt, 2001; Charney, 1984). Therefore, this study attempts to hold a small and informal workshop to strike a balance between classroom teaches' limited time schedule and the need to find a good way to increase rater consistency. By mainly extracting the part that a traditional rater training workshop or program usually has, i.e. a group discussion activity, in which raters can negotiate about their scoring criteria and the way they score students and thus try to reach a consensus among raters. In the study, this group discussion activity was called consensus building exercise.

Statement of the Problems

Many researchers have provided long term, large scale training, or a training session which includes many stages or meetings of raters (Luoma, 2005; Hughes, 2003; Underhill, 1991). However, researchers have argued that such long-term rater training has its disadvantages. For instance, Hamilton et al. (2001) reported that having all forty-five of their staff meeting at the same time during the busy exam period is very difficult. Charney (1984) argued that large rater training can be intimidating and might therefore have limited efficacy. In short, it is not easy for raters to attend regular training if it lasts several days, especially when they are working in their normal jobs. The problem of the impracticality of a traditional rater training can be solved through using a small workshop, like the consensus building exercise of the study.

While rater training has been emphasized by many researchers, most of the issues were focused on the effect of training on raters of compositions (Brown, 2002; Shohamy, Gordon, & Kraemer, 1992; Weigle, 1994, 1998). Fewer studies have examined the effect of rater training on speaking assessments, especially by analyzing both the intra-rater and inter-rater reliability. Further, little research has been done in EFL settings, such as Taiwan. Because of the limited number of related research studies about speaking assessment in EFL settings, there is a need to investigate whether a short-term workshop, like consensus building exercise, can truly improve intra-rater and inter-rater reliability of speaking assessment in university setting in Taiwan.

Purpose of the Study

Sims (2005) examined teacher training workshop for helping to make teachers' scoring to be more reliable. The result of Sims' (2005) study showed that those teachers gave relatively close scores to the same students after participating in the training workshop. However, the statistical analysis of teachers' intra-rater and inter-rater consistency and a further discussion of teachers' opinions about the training workshop were not investigated in Sims (2005).

Therefore, as an extension of Sims' study (2005), this study investigated the effect of the consensus building exercise on raters by examining raters' inter and intra-reliability before and after the consensus building exercise. By analyzing the change of intra-rater and inter-rater reliability, the researcher hopes to prove the applicability of such consensus building exercise. In addition, the researcher also interviewed the raters after the consensus building exercise in order to get more information from them; the analysis of the interview can be used a way to analyze raters' thoughts and opinions about the consensus building exercise. The researcher also attempts to prove that the consensus building exercise can not only provide practicability to suit classroom teachers' needs but also ease the worries of lower rater reliability of general human scoring in speaking test. In short, the purposes of this study are: 1) to investigate the effect of the consensus building exercise by analyzing intra-rater and inter-rater reliability, 2) to trigger raters' opinions about such kind of the short-term consensus building exercise through the interview with raters, and 3) to encourage the use of a small workshop like holding the consensus building exercise.

Significance of the Study

Through this study, the researcher hopes to make the following contributions to the field of English teaching and learning. First, the study was conducted to explore the possibility of implementing a short-term workshop like the consensus building exercise by examining the inter-rater and intra-rater consistency of the teachers. The researcher in inclined to prove that teachers will be able to spend less time being trained but still can improve their rater reliability in such short-term workshop. Second, the study can shed some light on the effect of rater training in a more detailed way not only by examining inter-rater and intra-rater consistency but also by discussing about the individual interviews with raters. Finally, the result may encourage classroom teachers to test students' speaking ability in order to achieve test validity without worrying about low rater reliability of speaking assessment or the limited time spent in attending a long-term training program. The researcher hopes this study will provide school teachers with practical suggestions of developing a more effective and efficient rater training workshop.

Research Questions

Three research questions will be addressed in this study:

1. Was there any difference in the agreement among raters' scores of video clips

before and after consensus building exercise? In short, what effects did the consensus building exercise have on inter-rater reliability?

2. To what extent an individual rater was able to consistently assign the same scores to each video clip? In other words, how did the intra-rater consistency of the scores change after the consensus building exercise?

3. What were the raters' opinions and suggestions about the consensus building exercise? What did the raters benefit from the exercise?

Definition of Terms

The following are brief definitions of terms used in this study:

Oral Interview: In the oral interview, the examinee interacts with the interviewer. Oral interviews are subjectively assessed by raters, and they may be assessed immediately or from an audio recording or videotape of the performance (Davies et al. 1999).

Rater reliability: Rater reliability includes intra-rater and inter-rater reliability (Bachman, 2005). Intra-rater reliability refers to the consistency within a rater himself and can also be called internal consistency. In other words, it is how individual raters can similarly apply their scoring criteria at different times or to different test takers. As for inter-rater reliability, it is the agreement among raters in their scoring. When a consensus or agreement among raters can be achieved, raters will avoid giving different scores (Alderson, Clapham & Wall, 2002; Bachham, 2005; Luoma, 2005).

Rating scales***:*** A rating scale is a series of short description of different levels of language ability. It describes how learners at different levels perform, thus helping raters to determine the level of test takers, resulting in more appropriate scores. However, it should be noted that too many level descriptors may sometimes distract raters' attention from scoring, so that raters cannot make a quick decision about Examinee's performance (McNamara, 2000; Underhill, 1991).

Rater training: Rater training is a series of activities held by educators or professional assessment developers. It provides more exposure to, and practice of, grading. During the training sessions, raters report their ratings and discuss the consensus of scoring. After this, the raters would discuss their opinions about how to modify the current rating criteria or scales. Therefore, rater training can not only maintain rater consistency but also generate more reliable and practical rating criteria or scales on the basis of different raters' opinions (Luoma, 2005).

Consensus Building Exercise: The exercise used in this study consists of three parts: group scoring, group discussion and group reports. Raters did group scoring in a group of two or there people and they tried to exchange their own ideas about the scores and tried to reach consensus within groups. Afterwards, each group of raters had group reports time to report their scores to the other groups; at this time, the raters negotiate the scores and discuss about the way they scored students.

CHAPTER TWO

REVIEW OF THE LITERATURE

With the increasing focus on communication in the classrooms, more and more teachers would like to integrate speaking activities into their classroom instruction. Researchers such as O'Malley and Pierce (1996) argued that one of the major responsibilities of language teachers is to enable learners to orally communicate with other people. Whether the instruction works in practice may be evaluated through assessment, which assumes that the assessment reflects the efficiency of instruction (Brown, 2004). Unfortunately, speaking assessment is quite challenging because it has many uncontrolled variables. For instance, Luoma (2005) indicated that speaking assessment is costly in time and money, and is often considered impractical. Brown and Yule (1999) also pointed out problems that may occur in human scoring. For example, raters may have poor understanding of, or be unable to hear, examinee's performance, and then give an inappropriate score. They may be biased or may interpret initial mistakes as the examinee's overall performance, or simply have idiosyncratic standards for performance. That is also why human scoring that calls for subjective judgment is often considered to have lower reliability. Hughes (2003) argued that whether a test is able to elicit appropriate behaviors or language functions that can effectively represent the examinee’s ability needs to be emphasized as well. Therefore, these problems may lead to unwillingness on the part of teachers test students’ speaking ability. Although speaking assessment is not easy, there is always a need to test students’ oral performance when considering the validity of a test. Brown (2001) defined test validity as whether the test really measures what it is intended to measure. In other words, if a teacher would like to test student’s speaking ability, the teacher should create a test that can reach validity. For example, oral interviews are more suitable than cloze tests to test students’ real speaking ability. Sims (2005) also stressed the importance of “the gains in validity of communicative oral testing, which outweighs the loss of objectivity” (p. 242). In short, a valid speaking test should actually drive students to open their mouths and use the language in an authentic and meaningful way (Brown, 2000).

This chapter consists of four sections related to speaking assessment. The first section of this chapter considers the nature of spoken language, such as pronunciation, grammar use, vocabulary use, and the relationship between speaking and listening skill. The second section covers important elements of speaking assessment, i.e. speaking tasks. The third section outlines different types of scoring and rating scales. The fourth section discusses the issue of rater reliability and then raises its implications for consensus building exercise and its effect on assessment.

The Nature of Speaking

This section consists of linguistic descriptions of spoken language, including the issue of pronunciation, grammar use, and vocabulary use in spoken language. Finally, the relationship between speaking competence and listening comprehension is discussed. These linguistic descriptions were also counted in the components of scoring criteria of this study.

The Issue of Pronunciation

According to Luoma (2005), people usually pay attention to what the speaker sounds like when he or she is speaking. The sound of one’s speech affects others’ impressions of one’s speaking ability. Luoma (2005) also observed that the sound of speech can refer to many features, such as separate sounds, pitch, stress or intonation. However, most people relate the sound of speech to pronunciation. When assessing learners’ speaking ability, the standard of judging pronunciation is quite controversial. As Brown and Yule (1983) and Morley (1991) have mentioned, it is difficult to determine a particular standard for foreign language pronunciation because people from different areas of English-speaking countries may have various accents or different ways of saying the same things, so that the standards cannot be identical. Further, some researchers argued that a learner's accent is quite acceptable in today's world (Leather & James, 1996; Pennington & Richards, 1986). They also agreed that a learners' accent not only represents an identity but also conveys a non-native status that enables others to avoid impoliteness. Luoma (2005) also mentioned that very few L2 learners can completely achieve a native-like standard, so the communicative effectiveness appears to be a more appropriate standard than pronunciation accuracy. However, choosing to test discrete elements in the language at one time, such as coping with sound recognition or choosing to focus on the communicative effectiveness relies on the purpose of the assessment (Underhill, 1991). Therefore, once the test designers define the purpose of the test, they can easily come out with scoring criteria based on the purpose.

Although pronunciation is a difficult issue to deal with, it still draws attention in speaking assessment for a variety of reasons. Luoma (2005) observed that deviation from native-like pronunciation is easy to notice, and thus convenient when giving grades. Additionally, pronunciation is often related to comprehensibility. Enabling people to understand what speakers are saying is essential for most L2 speakers because it could be a stepping stone to the beginning of communication. Therefore, the element of pronunciation should be included in the scoring criteria of speaking assessments.

Grammar Use in Spoken Language

Spoken language has its uniqueness. According to Chafe (1985), speech usually consists of shorter and simpler idea units, which may be illegitimate in writing. This is because speakers have to communicate their ideas with listeners in real time, long units which convey too many messages may confuse listeners (Luoma, 2005). Brown (2001) also mentioned that some speaking situations are more formal in which speakers try to use more literate grammar, such as speech, conference presentations, and expert discussions. Compared with written language, formal spoken language is still less formal than written language.

In many cases, the switch of word order is allowed in spoken grammar. Topicalisation and tails (McCarthy & Carter, 1995) are two examples that show how different word orders can be accepted in spoken language. McCarthy and Carter (1995) found that topicalisation, or thematic fronting, emphasizes the initial element of a clause, as in *That house in the corner, is that where you live?* By contrast, tails are noun phrases that come at the end of a clause, as in *he is quite a comic, that fellow.* These kinds of skills add uniqueness to the spoken language and thereby create naturalness. Luoma (2005) thus suggested that if the examinees can use these skills successfully and appropriately, they deserve reward for it.

Vocabulary Use

In addition to grammar and word order, vocabulary use in spoken language is also quite unique. According to Luoma (2005), speakers usually use lots of "generic or vague words", "fillers" or "hesitation markers" and "fixed phrases" in daily conversations (p.17-18). Generic words, such as *this one/that one*, and vague words, such as *thing* or *whatchamacallit,* are considered legitimate in spoken language but not in written language. Because speakers usually talk about people, things or activities that are familiar to them, even if these words are not precise, the speakers and the listeners can still understand each other. Moreover, these words work very well to make conversation more rapid, easy, and natural (Luoma, 2005). Another function of vague words in speaking proposed by Channell (1994) is that vague words help the speaker "hold the floor" of conversation regardless of missing words, and these words also attract listeners' attention to understand and further supply them if they can.

Speakers also make use of fillers or hesitation markers, like *ah*, *you know, you see, kind of*, etc. to create more time in which to speak. Hasselgren (1998) found that the more filler words learners used, the more they were considered as high-level by raters. However, Luoma (2005) pointed out that inappropriate use of fillers in a speaking test is sometimes considered a marker of non-fluency by raters, and that is also the difficulty of judging for raters. Additionally, fixed phrases, as Pawley and Syder (1983) and Nattinger and DeCarrico (1992) have explained, are often used by a speaker. Fixed phrases are expressions which come from speakers' mind automatically when a relevant situation happens. Sentences like *I'm doing all right* or *I thought you'd never ask* are prefabricated and thereby create much time for speakers to respond. ESL or EFL speakers usually memorize such kind of fixed phrase. Even though their ability has not reached at a certain level, they might be placed at the level which is prior to their real level. Therefore, Towell et al. (1996) indicated that listeners tend to interpret speakers who use more fixed phrases as having a higher level of linguistic competence.

Speaking Competence and Listening Comprehension

O'Mallery and Pierce (1996) considered speaking as a meaningful interaction among interlocutors because it usually involves more than one person. The interlocutors take turns together to hold the floor of conversation. O'Mallery and Pierce (1996) argued that speaking also means "negotiating intended meanings and adjusting one's speech to produce the desired effect on the listener" (p. 59).

Therefore, speaking skills have been considered to be closely interwoven with listening skills (Brow & Yule, 2001; Brown, 2004; Mohani & Mohtar, 2005; Murphy; 1991). According to Brown and Yule (2001), listening comprehension is actually the “process of arriving at a reasonable interpretation” of the speaker’s intended meaning (p.57). Brown (2004) also claimed that it is difficult to isolate speaking tasks from listening comprehension. Students as speaking test examinees may not comprehend what is heard and therefore cannot respond correctly (Mohani & Mohtar, 2005). How fluent a speaker can be sometimes depends on how well he or she can interpret the messages given by other speakers.

The close relationship between listening comprehension and speaking competence is also a headache for test designers because it is sometime hard to differentiate the mistakes a test taker made is due to poor listening comprehension or poor speaking ability. Murphy (1991) believed that listening and speaking are interdependent oral language processes and need to be taught and assessed in an integrated manner. Brown (2004) claimed that the test designer of an oral production test is challenged “to tease apart, as much as possible, the factors accounted for by aural intake” (p. 140).

Understanding the nature of spoken language can help test designers to create a valid test. Additionally, speaking tasks are also important elements that compose a test. The speech of examinees is be guided by different tasks and their language use also differ as purposes and contexts change. In the following section, speaking tasks is discussed in greater detail.

Speaking Tasks

O'Malley and Pierce (1996) stated that planning for speaking assessment involves identifying what kinds of tasks can be used. In this section, a brief introduction to speaking tasks is provided first, and then several factors that need to be taken into consideration when designing a task are presented. Finally, different task types are discussed, and oral interviews are given emphasis because this study uses oral interviews as a way to elicit the examinee's oral production.

Introduction of Tasks

Many researchers have provided their definition of tasks. Davis, Brown, and Elder et al. (1999) defined task as what a test taker is required to do during a test, such as participating in an interview or role play. In language learning contexts, tasks are also called language use (Luoma, 2005). Bachman and Palmer (1996) synthesized discussions by a number of applied linguists and found a general agreement among them: language tasks are 1) closely associated with specific situations, 2) goal-oriented, and 3) involve the active participation of language users. Bachman and Palmer (1996) concluded that speaking tasks are activities that involve speakers in using language to achieve particular goals or objectives in a particular speaking situation. They also believed that assessment of spoken language is most effective when it is preceded based on tasks. Some researchers have recognized that the effect of using a variety of tasks is the ability to make assessment more authentic and reliable (Hughes, 2003; Underhill, 1991). Since tasks play such an important role in speaking assessment, test developers need to be careful when designing a task.

Researchers indicated that test developers have to consider several things during the process of designing a task. For instance, Nunan (1993) proposed that the elements the task designer should pay attention to are input (materials that learners are to work on, such as radio broadcast and road map), goals, roles, and settings. Richards (1983) also indicated that content validity (is the task been used as part of instruction?), task validity (does memory play a role and influence the efficiency of the task?), purposefulness and transferability (can the purpose of the task can be transferred to real life?), authenticity (to what extent does the task measure actual spoken language?) are important elements which need to be considered when designing oral language assessment. Brown (2004) pointed out several other aspects that need to be accounted for. For example, he believed that speaking tasks better not to consist of only one skill, but should integrate other skills, such as listening or reading comprehension. In addition, the criterion that test developers have designed for a task is tricky because the description of criteria sometimes may not thoroughly contain the variability and comprehensiveness of oral production. Finally, it is also important to specify scoring procedures of the assessment so that a high reliability can be achieved, and examinees will have a better understanding about what is to be evaluated. In other words, specifying speaking assessment means to "define what kind of speaking will be assessed, how this will be done, and which aspects of the performances are going to be evaluated" (Luoma, 2005, p. 137).

Types of Speaking Tasks

Many researchers offer examples of speaking tasks that can be applied in an assessment setting (Brown & Yule, 2001; Brown, 2004; Luoma, 2005; O'Malley & Pierce, 1996; Underhill, 1991). By synthesizing their opinions, several common-used speaking tasks, such as oral interviews, picture-cued descriptions, radio broadcasts, video clips, information gap, story telling, role plays, oral reports, debates, etc. were suggested. Some of these researchers have grouped the speaking tasks in a systematic way. For instance, Brown (2004) provided taxonomy for speaking tasks. He grouped speaking tasks based on four categories, including imitative (repeating a word, a phrase, or a sentence), intensive (producing a short stretch of discourse through cued actions), responsive (having brief interaction an interlocutor by using short conversation), and interactive (producing long stretch of interactive discourse). O'Malley and Pierce (1996) also created an activity matrix especially for designing speaking assessment. Under each type of speaking task, they listed format (i.e. number of people involved), level of language proficiency, student preparation, and language functions to inform test takers what should be done in different tasks.

In this study, one of the speaking tasks, oral interview, was used to collect the materials. The oral interviews used in this study are designed for classroom application rather than large scale testing in order to elicit the examinee's oral production. There are many reasons for using oral interviews to elicit examinee abilities. For instance, Hughes (2003) and Underhill (1991) also indicated that the most commonly-used format for tests of oral interaction is the interview. Davis et al. (1999) argued that the function of oral interview tasks is for interviewers to effectively uncover the interviewees' attributes, experience or ability. Underhill (1991) mentioned that although oral interviews follow pre-determined structures, they still allow interviewees a degree of freedom to say what they authentically think. However, according to Hughes (2003), it should be noted that the examinee is usually passive and "unwilling to take the initiative" (p. 119). Accordingly, the interviewer may only elicit one style of speech lacking in a range of language functions, such as asking for information from the examinee (Hughes, 2003).

In conclusion, first, there is a general agreement among researchers that language tasks are specifically situated, goal-oriented, and involving the active participation of language users. Second, there are many concerns that a test designer must consider before creating speaking tasks for the assessment. For example, Nunan (1993) mentioned that a task designer should not only pay attention to materials that learners are using but also determine the goals, roles and settings when choosing which task will be involved in the test. Richards (1983) indicated the importance of content validity, task validity, purposefulness and transferability when designing an oral language assessment. Brown (2004) recommended that speaking should be assessed by integrating with other skills, such as listening skills. Furthermore, the grading criteria should be carefully designed. Finally, task types were also briefly reviewed, but oral interview was emphasized because it was used to collect material prepared for the study.

Another issue which has drawn attention from raters is the construction of the grading scale. A convincing and carefully-designed scale can also bring higher rater reliability because raters can consistently make a good use of it. Some of the most widely-used speaking scales are presented below.

Types of Scoring and Rating Scales

This part first discusses two types of scoring; objective and subjective scoring. Second, a comparison between holistic scoring and analytic scoring is made. Third, examples of different rating scales related to speaking assessment are reviewed, including the scale proposed by the American Council for the Teaching of Foreign Language (ACTFL), the scale of Test of Spoken English (TSE), and the scale of Common European Framework of Reference (CEFR).

Types of Scoring

There are two types of test scoring or marking, including objective and subjective (Alderson, Clapham, & Wall, 2002; Hughes, 2003). Objective scoring does not require scorers' personal judgment because the answer is usually either right or wrong. According to Sims (2005), objective scoring is used for tests like multiple-choice, true or false questions, and other item types for which the examinee can "produce a response which can be marked as either correct or incorrect" (p. 241). Objective tests are more reliable because if a rater can count the score properly, the test is said to have high both intra-rater reliability and inter-rater reliability.

While objective scoring is normally free from raters' opinions, subjective scoring requires raters to make judgments which are "more complicated than the right or wrong decision…their job is to assess how well an examinee completes a given tasks" (Alderson, Clapham, &Wall, 2002, p. 107). According to Henning (1987), Underhill (1991), and Hughes (2003), subjective scoring is usually used in writing or speaking assessment. It is said that the less subjectivity the scoring has, the greater the reliability it has (Cohen, 1994; Tuckman, 1988). Genesee and Upshur (1999) and Hughes (2003) indicated that writing or speaking assessments that use subjective scoring do not always have high reliability. However, Sims (2005) argued that there are ways of obtaining reliable subjective scoring for tests of speaking, that is, the need for a well-designed rating scale and the complement of rater training.

Rating Scales

Alderson, Clapham and Wall (2002) found that a well-designed rating scale enables a significant increase in the reliability of speaking assessment. According to Underhill (1991), a scale may consist of descriptions that explain "what the typical learner at each level can do" (p. 98). In general, scoring scales may consist of numbers, letters or other labels (e.g. excellent or good), which may be accompanied by a statement of behaviors that each point on the scale can be referred to (Alderson, Clapham & Wall, 2002; Davis et al., 1999).

There are two kinds of scale, holistic and analytic (McNamara, 2000). Davis et al (1999) called holistic scoring as “global assessment” or “impressionistic assessment” (p. 75). In holistic scoring, scorers usually grade examinee’s performance in its entirety. The raters are asked to pay more attention to overall communicative effectiveness, but not to particular aspects of the examinee’s production (Sims, 2005).

On the other hand, analytic scoring usually asks scores to individually judge several components of a performance separately, such as accuracy, fluency or pronunciation (Davies, Brown, & Elder et al., 1999). This type of scoring requires descriptors for each component. Hughes (2003) argues that analytic scoring has several advantages. First, analytic scoring solves the problem of “uneven development of subskills in individuals” (p. 102) and helps raters to consider more aspects of the performance and thus do not neglect these aspects easily. Third, the more subscores that a rater can give to the performance, the greater the reliability of the scoring. Finally, Brown (2001) stated that “the careful specification of an analytical scoring instrument can increase scorer reliability” (p. 387). In short, analytic scoring provides detailed guidance to raters, and offers rich information about the weakness or strengths of the examinee’s performance so as to increase rater reliability (Luoma, 2005). This study advocates the use of analytic scales for oral interview tasks. The components and descriptors will be presented later.

Examples of Speaking Scales

An example is the scale proposed by the American Council for the Teaching of Foreign Language (ACTFL, 1999). The ACTFL Speaking scale (ACTFL, 1999) is designed to investigate test-takers' functional competence, that is the ability to "accomplish linguistic tasks representing a variety of levels" (p.1). The test based on the scale is the called the Oral Proficiency Interview (OPI). This scale is holistic. It has ten levels, including superior; advanced-high, mid, low; intermediate-high, mid, low; novice-high, mid, low. Brindley (1998) mentioned that the ACTFL scale is a behavioral rating scale because the descriptions of each level indicate how learners use their language in specific contexts. In other words, speakers at each level are characterized by the ability to accomplish things listed in the description. The description of each level contains strengths and weaknesses that language learners who reach this level may have (See Appendix A).

Another example of speaking scales is the Test of Spoken English (TSE) Scale (ETS, 2007). TSE is a 20-minute audio-taped test of oral language ability. It is used to measure the ability of nonnative speakers of English to communicate effectively. The tasks in TSE are designed to elicit oral production in various discourses rather in separate linguistic components, such as pronunciation, grammar, or vocabulary use. There is no pass or fail score, score users can determine their own threshold. According to the TSE score user guide (2002), the TSE scale has five levels labeled from 20 to 60; each level describes the examinees' functional competence, sociolinguistic competence, discourse competence and linguistic competence. Test takers of TSE will be asked to tell a story, describe a graph, and answer questions. Their answer will be recorded and scored by trained raters. (See Appendix B and C)

The Common European Framework of Reference (CEFR) (Council of Europe, 2001) also provided a speaking scale. According to Council of Europe (2001), this scale is used as a basis for creating test-specific criteria rather than developing for any particular test. Luoma (2005) suggested that teachers can analyze students' performance by referring to the scale to see if learners' performances correspond to the descriptions at each level. This scale has six levels, labeled as A1/A2 (basic), B1/B2 (independent), C1/C2 (proficient). The criteria of this scale are analytic because it covers five linguistic components, including range, accuracy, fluency, interaction and coherence. Each level descriptor contains several statements which describe what the learners should do at each level. (See Appendix D)

Rater Reliability and Rater Training

Luoma (2005) pointed out that the reliability of speaking scores depends on high-quality scoring instruments and procedures. A number of factors may influence raters' impressions of examinees' oral proficiency, such as interviewers' behaviors and their nationalities, which may lead to low rater reliability (Brown; 2003, Chalhoub-Deville & Wigglesworth, 2005; Lumley & McNamara, 1995; Nakatsuhara, 2007). Many researchers have proposed rater training as a way to reach rater reliability (Brown, 2003; Congdon & McQueen, 2000; Elder, Knoch, Barkhuizen, & Randow, 2005; Lumley & McNamara, 1995; Nakatsuhara, 2007; Weigle, 1998). In this section, factors that affect raters' impressions of how well a person can speak a language are discussed first. Most of these findings have implications for training programs, so the effect of rater training is also discussed.

The Influence of Rater or Interviewer Factors on Speaking Assessment

It is difficult to avoid the effects of other variables that may affect the assessment. For example, Nakatsuhara (2007) investigated how the raters' impressions of the examinee's ability were influenced by two different interviewers' behaviors in an oral interview test. The examinees were generally scored higher in the categories of 'pronunciation' and 'fluency' when being interviewed by Interviewer B whose behavior was more non-test-like and had less control of the interaction. Nakatsuhara (2007) believed that is because the examinee can talk with greater freedom in directing the conversation; hence and their fluency in the conversation increased. In addition, because the examinees had superiority in speaking, they made a good use of avoidance strategies, avoiding certain words whose pronunciations they were not sure of, leading to higher scores in pronunciation (Nakatsuhara, 2007). In the interview with Interviewer A who usually stated his questions explicitly and had a higher control of the topic development, the examinee only got a higher score in the category of 'vocabulary.' According to Nakatsuhara (2007), this finding resulted from the teacher-like behaviors of interviewer A, which may "push the examinee to her limits of vocabulary resources" (p.8).

Similarly, Brown (2003) investigated interviewer variation, focusing on how the two interviewers utilized different communicative strategies in order to continue the conversation, hence influencing both the examinee's performance and raters' grading. Her study showed that these two interviewers did have distinct individual behaviors, which resulted in a great impact on the examinee's performance and also raters' impressions of examinee proficiency. The examinee got higher scores in the interview with Interviewer C, who asked more open-ended questions, illustrated her questions in a more explicit way, was more consistent in developing topics, and was better at breaking down unsuccessful prompts into a more explicitly focused question than the other interviewer, and gave more positive feedback by acknowledging the examinee's opinions or by asking extended questions to show her understanding. In the interview with Interviewer C, the examinee can successfully hide his interpretation of the pragmatic force of the prompts. For example, the examinee sometimes viewed interviewers' extended questions as only confirmation requests, leading to long silences during the interaction due to this wrong interpretation, which in turn may result in raters' impression of poor ability. This poor proficiency was successfully "hidden" in the interview with Interviewer C.

Both Brown (2003) and Nakatsuhara's (2007) studies addressed interviewer's training. Not only is rater training is needed, but interviewers require training for an interaction-based test. Sometimes, the low reliability of an oral test is due to interviewer variations, not raters alone. Because interviewer variation may affect the test-taker's score, the interviewers' behaviors have to be taken into consideration when creating the test. In a classroom setting, a teacher often plays two roles at the same time, an interviewer and also a rater. Thus, interviewer training is as necessary as rater training.

As the above discussion have mentioned, an interviewer who successfully provides support that "directs the attention of the learner to key features of the environment, and which prompts them through successive steps of a problem" (Mitchell & Myles, 2004, p.195) can facilitate test-takers' speaking fluency and help test takers scaffold their confidence. During the process of “scaffolding” (Wood, Bruner, & Ross, 1976, as cited in Mitchell & Myles, 2004, p.195), test-takers’ “affective filter” (Krashen, 1982, as cited in Mitchell & Myles, 2004, p. 48) would become lower because the load on speaking may be partially shared by interviewers, or the conversation flow can be efficiently guided by them.

Sometimes, raters’ personal factor may also interfere with the score result, such as age, experience, or nationalities of raters. For example, Chalhoub-Deville and Wigglesworth, in 2005, examined how raters’ nationalities affected their judgment of speaking proficiency. In this study, participants included 12 test-takers and 104 raters who were divided into four groups based on their nationalities. Raters were also interviewers at the same time, so they had to grade the test-takers while interviewing them. The test-takers were asked to respond to 12 audio-taped tasks grouped into three different types, including give and support an opinion, picture-based narration and presentation (Chalhoub-Deville & Wigglesworth, 2005). The results showed that indeed there were differences among raters when scoring the three different tasks, but only one percent of the variance in the task measures can be attributed to the influence of rater’s nationalities. Therefore, the researchers also implied that there might be other variables which influenced their scoring as well. One limitation of this study is that the researchers only analyzed test-takers’ proficiency by evaluating their performance in only three tasks. Results may be different if the participants are examined in terms of other detailed scales that have several categories used to specifically describe test-takers' oral proficiency, such as grammar and pronunciation.

Effects of Rater Training

Although Brown (2003) and Nakatsuhara's (2007) suggested that interviewer training is important, a training program is usually held for raters because they are the people who give scores. Rater training plays an important role in achieving raters self consistency, i.e. intra-rater reliability, and peer agreement among different raters, i.e. inter-rater reliability. Therefore, several studies have investigated the effect of rater training (Charney, 1984; Congdon & McQueen, 2000; Cook, 1989; Elder et al., 2005; Hamilton et al., 2001; Weigle, 1994, 1998).

Large-scale assessment is usually considered as a high-stakes test, for it has a great influence on the future use of the score, such as student placement or job pursuit. Therefore, it is essential to maintain high rater reliability in a large-scale assessment in order to achieve test fairness. Congdon and McQueen (2000) investigated ten raters' stability and mutual agreement before and after a training program. They found that there was little effect on absolute agreement among raters, that is, inter-rater reliability, because each of them held different opinions of the ways to grade test-takers. Weigle (1994, 1998) also conducted two studies investigating difference between PRE (before training) data and POST (after training) data in writing assessment. She summarized her finding and mentioned that rater training was "more successful in helping raters give more predictable scores (i.e., intra-rater reliability) than in getting them to give identical scores (i.e., inter-rater reliability)" (p. 263). Weigle (1994, 1998) explored the effect of training on writing assessment, finding that an overemphasis on achieving close scores may sacrifice raters' expertise and experience or even the basic interaction between text and reader.

Although some researchers have indicated rater training did not work very well in increasing inter-rater reliability (Congdon & McQueen, 2000; Weigle, 1994, 1998), some researcher believed inter-rater reliability may still be achieved through sufficient feedback provided by peer raters and through follow-up training (Congdon & McQueen, 2000; Elder et al., 2005). Elder et al. (2005) argued that the individual feedback given by each rater during the training session can not only enhance the self consistency of individual raters but also reduce rater diversity among raters, i.e. increase inter-rater reliability. Elder et al. (2005) made a comparison before and after raters gave feedback to each other. The results showed that most of the raters found feedback useful in facilitating their awareness of rating behaviors. By receiving feedback from their peers, they could modify their rating and make sure that they were on the right track in scoring. The implication of this study is that peer feedback is beneficial to enhancing inter-rater reliability. Not only does it increase inter-rater reliability, but it can be helpful in improving intra-rater reliability as well. It allows raters to have self-examination of their own scoring and compare their ideas with other raters. Elder et al. (2005) suggested that there is a need to include discussion sessions to permit raters to give feedback to each other in a training program. In addition, Congdon and McQueen, (2000) also indicated the need for ongoing training during the rating period in order to increase the effectiveness of rater training. Except for a small deficiency of training, rater training is considered to have many positive effects on the raters, especially in helping raters to have better comprehension of the whole rating scale, and the improvement of individual raters' self-consistency.

There are some more researchers approved the effectiveness of rater training in both increasing intra-rater and inter-rater reliability. For example, Cook (1989) pointed out that rater training effectively reduces not only extreme leniency or harshness but also the halo effect in speaking assessment. The halo effect is generally defined as a bias in which the estimation of one characteristic of a person affects the estimation of another characteristic of that person (Spear, 1996). In other words, raters might sometimes expect higher scores from certain students, such as students who have higher academic achievement in their classroom performance. However, Cook (1989) claimed rater training can reduce the halo effect and reduce the extreme scoring in order to increase fairness.

There are some issues about rater training that need to be concerned. Charney (1984) observed that raters usually vary in time they need to judge students' performances, and that small scale training is more applicable in classroom settings. Hamilton et al. (2001) raised the problem of impracticability of rater training, that is, time issue. Rater training generally has a scheduled syllus, but it sometimes becomes a burden for classroom teachers. Hamilton et al. (2001) also pointed out the limitation of implementing large scale rater training program and therefore suggested a method that allows more flexibility for raters, on-line rater training. However, on-line training still has limitations, such as the teachers not using the site due to either the lack of time or a dislike of reading materials online. Further, the efficacy of on-line training is affected by familiarity with computers and Internet navigation. If raters are unfamiliar with those facilities, the training may fail to be effective (Hamilton et al., 2001).

In conclusion, this section analyzed the impact of interviewers' nationalities and the influence of their behaviors on raters' giving scores, and addressed the importance of interviewer training. Rater training is as important as interviewer training, so the effect of rater training was reviewed. Although rater training helps to achieve high rater reliability, it also has some disadvantages, such as impracticality and limited time. Even if on-line training is used to allow more flexibility for raters, the preference for using reading materials rather than looking at the screen, and the issue of familiarity with computer facilities, creates problems for the use of on-line training. Therefore, a short term rater training may have its applicability and superiority in classroom settings.

CHAPTER THREE

METHODOLOGY

This chapter describes the research method used in this study. The research method includes four sections: the participants, the materials, the procedures of data collection, and the procedures of data analysis. In this study, ten teachers from the Foreign Languages and Literature Department (FLLD) of Tunghai University were the participants. They attended a consensus building exercise and scored some video clips of students being interviewed. The consensus building exercise intended to help the teachers to increase their internal consistency (intra-rater consistency) and to reach agreement among raters (inter-rater consistency) of their scores for video clips. The material consisted of three parts: the grading criteria and guidelines, the video clips of students being interviewed by an experienced teacher from the FLLD, and the consensus building exercise and its follow-up activities which was hold after one month. The data for the study was collected from the teachers' scores for the video clips before and after the consensus building exercise. Finally, statistical procedures of Pearson Product-Moment Correlations (γ) were used to analyze teachers' inter-rater consistency and the calculation of point difference between teachers' first scorings and the second scorings of the same set of video clips were used to investigate teachers' intra-rater consistency.

Participants

The participants were ten teachers (seven females and three males) from the Freshman English for Non-majors (FENM) program in the FLLD of Tunghai University. Because the goal of the FENM program intends to provide the opportunities for students to practice their four skills (speaking, listening, reading and writing), teachers of FENM program did really have many chances to teach and assess students' oral proficiency. In the study, teachers' number of years in the FENM program is also an indication of teachers' experiences of assessing students' oral performance. Therefore, five of the participants were more experienced teachers who have been teaching for more than five years, while the other five of the participants were new teachers who have been teaching for less than five years. All of the teachers had masters or doctorates in TESOL or a relevant field. Two of them were native speakers, and the other eight were native-like speakers of English. The video clips of students being interviewed were labeled as S1-S12. The raters were the teachers who scored the video clips. The new teacher raters were labeled as N1-N5 (N equals NEW teachers), and the experienced teacher raters were labeled as

E1-E5 (E equals EXPERIENCED teachers).

Materials

The materials that the study used are as follow: first, the grading criteria and guidelines; second, teachers' scorings of the twelve video clips in which students had been interviewed; third, the consensus building exercise and its follow-up activity one month after the exercise and fourth, the interview with the raters.

The Grading Criteria and Grading Guidelines

Grading criteria and grading guidelines not only helped to select students but also were used by the raters in the consensus building exercise when they were giving scores to the video clips. The grading criteria the raters used to evaluate students' overall oral proficiency were based on a thirty point scale. The total score for the interview is thirty points, which reflects the policy of the FENM program at Tunghai University; in other words, this kind of thirty point scale is actually used by the teachers in their FENM classroom.

The criteria for the guidelines were adopted from Harris (1986), Hennings (1987), Hughes (2003), and Sims (2005). The scale was composed of five components: 1) content, 2) accuracy, 3) fluency, and 4) pronunciation and 5) vocabulary (See

Appendix E and F). According the requirement of the FENM program at Tunghai, oral assessment takes thirty percent of the total of both midterm and final exams. As a result, the criteria used in this study were a thirty-point scale which is applicable to the FENM program. The proportion of each component was also discussed by the FENM teachers for the need of their classroom evaluation.

The content component was worth ten points. It was used to see if student performances were knowledgeable, substantive, thoroughly responsive, and relevant. It implied raters have to ask themselves the following questions: a) *did the student answer the question*, b) *did what the student says make sense*, and c) *was it relevant to the topic?*

The accuracy component was worth five points. It was used to see if students' performances included appropriate tense use, word order, pronouns, and complete sentences. In other words, the accuracy component was used to test examinee grammar production. The fluency component was also worth five points. This component was used to see if the students had coherent language and confidence language use. The fluency component also includes the speed of oral delivery.

The pronunciation component was worth five points. Raters asked themselves questions like "*was the students' pronunciation clear?*" or "*could what the student says be understood?*" The vocabulary component was also worth five points. In this category, raters judged if students had appropriate register, effective word choice and usage, and a sophisticated range in vocabulary. In short, questions like "*was the vocabulary appropriate and used correctly*", and "*was there a range in vocabulary?*" ("*did the student use more than just a few words?*) might have been answered by raters when they were rating students.

The Collection of Students' Video Clips

Twelve video clips in which twelve freshmen had been interviewed by one experienced teacher who has been teaching for more than fifteen years in the FENM program were prepared. These videotapes were labeled from S1 to 12.

During the interviews, the interview questions were arranged in the following stages as recommended by Canale (1984) and Underhill (1999). There were three main stages: introduction and warm-up, find level, and check level. The first stage, introduction and warm-up, was to help the interviewees to become more comfortable and less anxious about the interview situation. The second stage, find level, was to help the interviewer to establish approximate level of the interviewees. The last stage, check level, was to confirm if the level establishment was right. Table 3.1 shows how each question fit into the stages.

Table 3.1

The Interview Stages and Interview Questions

Introduction and warm-up	Q1. What is your name? Q2. What is your student number?
Find level	Q3. What is your favorite activity? Q4. Tell me about your major and about your studies? Q5. What language other than your native language that you most likely to learn? Q6. Name your worse habit. Q7. What would you do if you suddenly won one million dollars?
Check level	1. Extended question of Q3 (e.g.) Where do you go swimming? How long do you jog? 2. Extended questions of Q4 (e.g.) Do you like it? Is it easy or difficult? What do you like about it? 3. Extended questions of Q5 (e.g.) Why do you want to learn this language? Have you

	ever been to the country before? 4. Extended question of Q6 (e.g.) How are you going to break this habit? 5. Extended question of Q7 (e.g.) Are you going to buy anything for yourself?

The interviewer asked many extended questions or sometimes paraphrased the questions during the interview. Take Question 4 for example, when the interviewee had a hard time answering Question 4 ‘*Tell me about your major and about your studies?*’ , the interviewer might paraphrase the question to ‘*Do you like it?*’ or ‘*Is it easy or difficult?*’ These were the elicitation techniques suggested by Hughes (2003), i.e. requests for elaboration and paraphrasing the questions. Elicitation techniques were commonly used during the stage of find level, not only for triggering more responses from the examinee, but also for checking and fine-tuning the perception of the examinee’s level.

Thirty students were selected from the FENM program. They were videotaped while they were being interviewed. These students were interviewed in room 101, located in the FLLD building. Each student was interviewed for about two to three minutes. An effort was made to include students representing three levels of oral proficiency, which are high, mid and low. Therefore, the interviewer holistically grouped these students into three levels: high, middle and low according to the criteria used in this study. Eventually, only twelve student video clips (four female and eight male students) were selected because these twelve students displayed a more obvious differentiation in their level of performance. For each level, there were an equal number of student video clips (four for each level). Table 3.2 shows how the twelve student video clips were arranged.

Table 3.2

The Arrangement of Video Clips

Level	1st set	2nd set
High	S3, S4	S10, S11
Mid	S5, S6	S7, S12
Low	S1, S2	S8, S9

The researcher divided the twelve video clips into two sets, six video clips for each set. During the consensus building exercise, the video clips were played in alphabetical order, so that the raters would not make hypotheses about the students' level when they were giving scores. Before beginning with the consensus building exercise, the raters scored the first six video clips (S1-S6). During the exercise, the raters worked together in groups to score the first six video clips (S1-S6) again with their group members. After the consensus building exercise, the raters were asked to watch another six video clips and gave scores to them (S7-S12). One month later, the raters watched the same twelve video clips again and scored them.

The Consensus Building Exercise and its Follow-up Activity after One Month

This study seeks to evaluate to what extent a raters' consensus building exercise can influence teachers' intra-rater reliability and inter-rater consistency in oral assessment. Therefore, the consensus building exercise was held to collect the data of the study.

The implementation of the whole process included five parts: 1) the presentation and explanation of the grading criteria and guidelines; 2) the first independent scoring before the exercise (to score the first set of video clips S1-S6); 3) the consensus building exercise: group scoring (to also score the first set of video clips S1-S6), group report, and group discussion 4) the second independent scoring (to score the second set of video clips S7-S12), and 5) the follow-up activity one month after the consensus building exercise (to score the same twelve video clips S1-S12).

The whole process began with the presentation and explanation of the grading criteria and guidelines by the researcher. The teachers tried to familiarize with the criteria and guidelines and raised their questions if there was any. They were asked to score these video clips by referring to the grading guidelines and using a grading sheet provided to them.

Before beginning with the consensus building exercise, the raters first practiced to score a set of video clips (S1-S6). Next, the teachers were divided into a group of two of three people. They started the consensus building exercise by first having a time of group scoring, which could provide a chance for teachers to exchange their opinions about scoring criteria and thus helped them reach a consensus on the scores for each video clip. Later on, the teachers also had to report their results to the other groups. This was also a time for one group of teachers to discuss about their results with the other groups. During this process, these teachers tried to express their opinions and negotiated their scoring criteria with the other groups of teachers, and thus obtained individual feedback from each other. As suggested by Elder et al. (2005), there is a need to for a rater to obtain individual feedback of his scores for the examinees from his peers in a rater training program because it helps to enhance the awareness of his own rating behavior and also helps to reach peer consensus. The whole exercise lasted about thirty minutes.

One month after the consensus building exercise, the raters came to meet again in a follow-up activity to watch the twelve videos which were the same video clips they have watched one month before. The one month period between the carryout of the consensus building exercise and its follow-up activity was an attempt to avoid teachers' impression of their previous scoring and to cause as less as possible influence on their second scoring. In this follow-up activity, the raters did not have any group activities, such as discussion, group scoring, and group report as previous occasion. All they had to do was to watch the same twelve video clips again and score the students independently.

The Interview with Raters

In order to have a thorough understanding of raters' thoughts and opinions about the consensus building exercise, the researcher individually interviewed these raters after they accomplished the follow-up activity of the exercise. The interviews with raters began about six months after they completed the follow-up activity, and the interviews were recorded throughout the whole process. The interview data can yield some interesting insights into raters' differences of the way they scored students' video clips, which can not be gotten from other statistical method.

Since the interviews happened about six months after the whole procedure, most of the raters' more or less lost their memories of things happened during the exercise and its follow-up activity. Therefore, a list of interview questions was mainly made to trigger the raters' memories about what they were doing, thinking or feeling during the exercise and the activity. Some of the interview questions were designed mainly based on Weigle's (1994) study, such as the questions about whether the teacher revise their expectations of students and thus changed their scoring criteria and the questions about whether the teachers had more concerns for rater agreement and self consistency after the exercise. The interview questions began with raters' background investigation first, and then some questions developed from Weigle's (1994) study were asked. Finally, the questions about raters' opinions and suggestions about the exercise or if there is any benefit they got from the exercise were listed. The interview question list is listed in Appendix G.

Data Collection Procedures

Data was collected from the three sources; first, raters' consensus building exercise; second, the follow-up activity one month after the exercise, and third, the interview with raters six month after the follow-up activity. In the consensus building exercise, three parts of procedures were included: group scoring, group reports on their score results, and group discussion of their scoring criteria and the score results. As for the follow-up activity one month after the exercise, the raters only did independent scoring of the same video clips they had watched during the exercise.

The consensus building exercise took place in room FL005 located in the FLLD building, on March 27, 2008. The exercise was videotaped throughout. Before the consensus building exercise, the raters were first familiarized with grading guidelines and grading criteria. Next, the raters watched six video clips (S1-S6) in which students were being interviewed and then worked independently to give scores to each video clip. This was the data of teachers' first scorings of S1-S6, and it was before the implementation of the exercise.

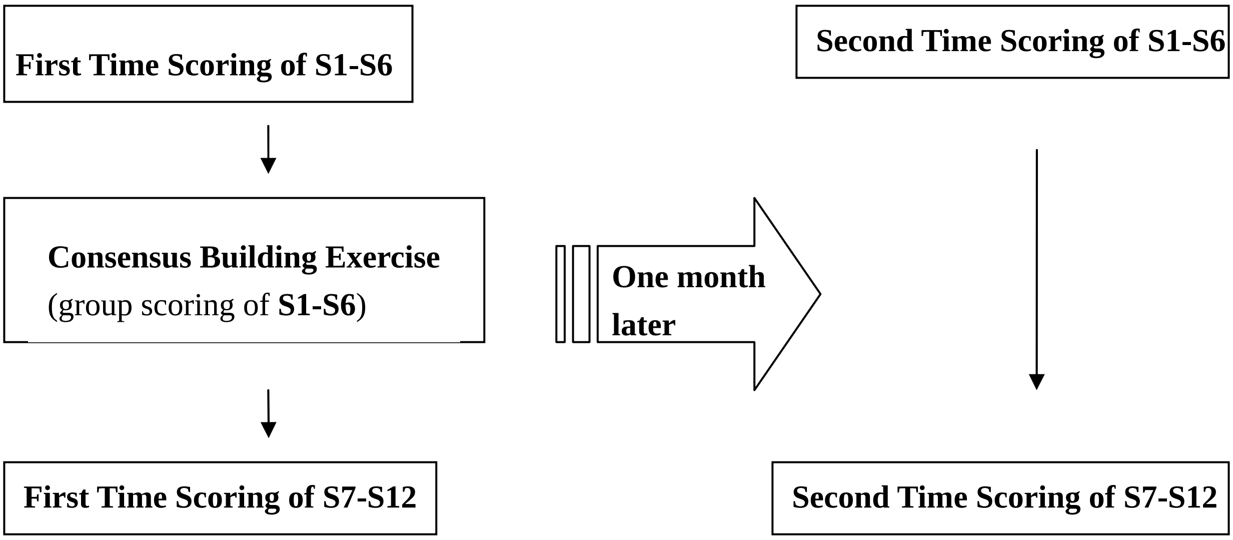
After the raters watched and scored the first six videos (S1-S6), they began the consensus building exercise with working together in groups of two or three people to discuss their previous scores for these video clips (S1-S6) and their scoring criteria. During the exercise, the teachers tried to come to a consensus of group scores for these six video clips. The result of teachers' group scorings was not included in this study. After the group scoring, each group reported their results to the other groups. There was a group discussion of all teachers in which every teacher made some comments about each group's score results and expressed his or her own scoring criteria for the video clips. Next, the raters watched another six video clips (S7-S12) and gave scores independently. This was the data of teachers' first scorings of

S7-S12. The collection of this data was used to see if the raters could apply the consensus of the scoring criteria they just achieved in the exercise when they scored different video clips.

After one month, the raters got together again and had a follow-up activity. They watched the same twelve video clips again and scored them independently. The raters did not have any group activity at this time. Two of the raters couldn't attend the follow-up activity, so they were asked to watch the video clips and give score to those video clips alone in their office. These data were the second scorings of S1-S6 and S7-S12 individually. The collection of the second scorings of S1-S12 was used to see if the effect of the exercise still last after one month. Figure 3.1 shows how the whole procedure was carried out.

Figure 3.1

The Consensus Building Exercise and its Follow-up Activities


First Time Scoring of S1-S6
Second Time Scoring of S1-S6
Consensus Building Exercise
(group scoring of S1-S6)
One month later
First Time Scoring of S7-S12
Second Time Scoring of S7-S12

Another data for this study is the interview with raters. By analyzing the interview with raters, the effect and the influence of the consensus building exercise of the study can also be investigated. An interview with the raters could be a tool to clarify their thoughts and opinions about the consensus building exercise, and it was also helpful to explain the result. During the interview, raters' thoughts were prompted and verbalized without being disturbed very often. The individual interviews with the teachers were done by the researchers after the follow-up activity. The duration of the finish of follow-up activity and the interviews with teachers was about six months. During the interview with teachers, the use of the interview questions were just to make teachers' responses relevant to the consensus building exercise and to trigger teachers' memories about the exercise as well. The interview questions are listed in Appendix G.

Data Analysis Procedures

The data analysis in this study explores the following questions: 1) Was there any difference in the agreement among raters' scores of video clips before and after consensus building exercise? In short, what effects did short-term consensus building exercise have on inter-rater reliability? 2) To what extent an individual rater was able to consistently assign the same scores to each video clip? In other words, how did the intra-rater consistency of the scores change after the consensus building exercise? 3) What were the raters' opinions and suggestions about the consensus building exercise? Did the raters benefit from the exercise?

According to Bachman (2005), in order to estimate their inter-rater consistency, Person Product-Movement Correlations (γ) between all pairs of raters were calculated. This approach only indicated which pairs of raters agree with each other the most: the higher the correlation was, the higher the inter-rater consistency would be. Ten raters yielded forty five correlations, and thus the average of these forty five correlations was calculated to give an estimate of the average agreement among the ten raters. There were four averages of correlations, including the scoring of S1-S6 at the first time and at the second time, and the scoring of S7-S12 at the first time and at the second time. These four correlation values were compared to see if the consensus building exercise had any influence on the teachers' inter-rater consistency.

As suggested by Weigle (1994), one way to investigate intra-rater consistency is through a comparison made of the scores given to the same video clips by the same raters from the first time and the second time. Thus, in order to answer the second research question, a comparison between raters' mean scores of the first scorings for video clips and the mean scores of the second scorings for the same video clips was made. This involved subtracting each individual rater's mean score of the second scorings from the mean score of the corresponding first scorings. The absolute value of this number is the point difference between each first and second scorings. The calculation procedure was done for each component and for the total scores. For this study, an extreme difference in points between the first scorings and the second scorings for content component was defined as a difference of greater than 1 point. For the other four components: accuracy, fluency, pronunciation and vocabulary, a difference of greater than .5 point was considered as an extreme case. As for the total score, a difference of greater than 3 points was an extreme difference in points between the first scorings and the second scorings.

In order to answer the third question, the researcher interviewed the ten raters individually after the follow-up activity had finished. The duration of time was about six months. The interview questions were mainly designed to have the raters talk about their opinions about the consensus building exercise, and how they benefited from the consensus building exercise. Because the interview with teachers happened about six month after the follow-up activity, the function of interview questions could also help the teacher pick up their memories about what they have done during the consensus building exercise and the follow-up activity.

CHAPTER FOUR

RESULT AND DISCUSSION

This chapter covers the following sections: 1) the effects of short-term consensus building exercise on inter-rater reliability; 2) the effects of the short-term consensus building exercise on intra-rater reliability and 3) the analysis of the data collected from the interviews of the raters.

The Effect of the Consensus Building Exercise on Inter-rater Consistency

In this section, the researcher's intent was to answer the first research question: *Was there any difference in the agreement among raters' scores of video clips before and after the consensus building exercise? In short, what effects did the consensus building exercise have on inter-rater reliability?*

In order to investigate ten teachers' inter-rater consistency of scoring the same set of video clips before the consensus building exercise and their inter-rater consistency after the exercise, the researcher first calculated Pearson Product-Moment Correlation coefficient (γ) between all pairs of raters, and then calculated the average of these correlations as an estimate of inter-rater consistency. This method of calculating inter-rater consistency was based on Bachman’s (2005) standard procedure of evaluating inter-rater reliability. The correlations between each pair of raters show which pair agree with each other the most, and the average of these correlations gives an indication of the average agreement among the ten raters. By comparing the averages of correlations before and after the consensus building exercise, whether the consensus building exercise had any influence on teachers’ inter-rater consistency could be revealed. In this study, ten raters yielded a set of forty-five correlations of each pair. Thus, the average of these forty-five correlations was estimated as an indication of inter-rater consistency. All raters’ scores of all video clips are presented in Appendix H-K.

Table 4.1 and Table 4.2 present the correlations of raters’ scorings of S1-S6 from the first time (before the consensus building exercise) and from the second time (one month after the exercise). Similarly, Table 4.3 and Table 4.4 show the correlations of raters’ scorings of S7-S12 the first time (immediately after the consensus building exercise) and the second time (one month after the exercise).

Table 4.1

Pearson Correlations of Raters' Scores for S1-S6 from the First Time

	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5
N1	1	.773	.814*	.917*	.837*	.688	.612	.855*	.929**	.727
N2	.773	1	.917**	.714	.863*	.805	.701	.982**	.870*	.618
N3	.814*	.917**	1	.792	.873*	.948**	.722	.948**	.956**	.584
N4	.917*	.714	.792	1	.768	.787	.843*	.758	.842*	.861*
N5	.837*	.863*	.873*	.768	1	.749	.660	.895*	.857*	.764
E1	.688	.805	.948**	.787	.749	1	.821*	.815*	.854*	.578
E2	.612	.701	.722	.843*	.660	.821*	1	.649	.618	.856*
E3	.855*	.982**	.948**	.758	.895*	.815*	.649	1	.941**	.603
E4	.929**	.870*	.956**	.842*	.857*	.854*	.618	.941**	1	.576
E5	.727	.618	.584	.861*	.764	.578	.856*	.603	.576	1

Note. * Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

Table 4.2

Pearson Correlations of Raters' Scores for S1-S6 from the Second Time

	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5
N1	1	.730	.939**	.918**	.786	.923**	.617	.920**	.994**	.877*
N2	.730	1	.757	.857*	.937**	.752	.799	.681	.744	.640
N3	.939**	.757	1	.981**	.794	.784	.796	.889*	.924**	.967**
N4	.918**	.857*	.981**	1	.889*	.787	.856*	.885*	.903*	.917**
N5	.786	.937**	.794	.889*	1	.720	.845*	.838*	.780	.684
E1	.923**	.752	.784	.787	.720	1	.410	.745	.936**	.670
E2	.617	.799	.796	.856*	.845*	.410	1	.736	.608	.790
E3	.920**	.681	.889*	.885*	.838*	.745	.736	1	.910*	.866*
E4	.994**	.744	.924**	.903*	.780	.936**	.608	.910*	1	.871*
E5	.877*	.640	.967**	.917**	.684	.670	.790	.866*	.871*	1

Note. * Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

Table 4.3

Pearson Correlations of Raters' Scores for S7-S12 from the First Time

	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5
N1	1	.879*	.981**	.946**	.818*	.899*	.870*	.855*	.967**	.864*
N2	.879*	1	.943**	.855*	.938**	.959**	.883*	.863*	.948**	.969**
N3	.981**	.943**	1	.945**	.914*	.938**	.870*	.908*	.990**	.920**
N4	.946**	.855*	.945**	1	.823*	.913*	.884*	.955**	.893*	.837*
N5	.818*	.938**	.914*	.823*	1	.884*	.735	.907*	.910*	.902*
E1	.899*	.959**	.938**	.913*	.884*	1	.966**	.891*	.929**	.982**
E2	.870*	.883*	.870*	.884*	.735	.966**	1	.800	.858*	.927**
E3	.855*)	.863*	.908*	.955**	.907*	.891*	.800	1	.850*	.828*
E4	.967**	.948**	.990**	.893*	.910*	.929**	.858*	.850*	1	.936**
E5	.864*	.969**	.920**	.837*	.902*	.982**	.927**	.828*	.936**	1

Note. * Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

Table 4.4

Pearson Correlations of Raters' Scores for S7-S12 from the Second Time

	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5
N1	1	.827*	.935**	.789	.724	.708	.906*	.870*	.867*	.779
N2	.827*	1	.956**	.801	.833*	.720	.927**	.832*	.836*	.891*
N3	.935**	.956**	1	.891*	.883*	.830*	.974**	.940**	.918**	.905*
N4	.789	.801	.891*	1	.872*	.974**	.920**	.935**	.971**	.942**
N5	.724	.833*	.883*	.872*	1	.914*	.875*	.954**	.784	.789
E1	.708	.720	.830*	.974**	.914*	1	.863*	.943**	.894*	.858*
E2	.906*	.927**	.974**	.920**	.875*	.863*	1	.942**	.939**	.903*
E3	.870*	.832*	.940**	.935**	.954**	.943**	.942**	1	.894*	.839*
E4	.867*	.836*	.918**	.971**	.784	.894*	.939**	.894*	1	.963**
E5	.779	.891*	.905*	.942**	.789	.858*	.903*	.839*	.963**	1

Note. * Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

As mentioned previously, ten raters generated forty-five correlations. Thus, Table 4.5 shows the average of these correlations from the four different sessions. The average of correlations gives an estimate of inter-rater consistency.

Table 4.5

The Averages of Correlations

	S1-S6	S7-S12
First Time	.79	.90
Second Time	.81	.88

As can be seen from Table 4.5, raters' inter-rater consistency of scorings of S1-S6 from first time was .79 which was very high. The first scorings of S7-S12 happened immediately after the exercise, and the correlation of it was higher (.9) even though the raters did not use S7-S12 during the exercise. The result implies that the raters might still bring a new idea of scoring students to their future scoring. During the exercise, the raters synthesized their criteria with the other raters' criteria and tried to reach a consensus among raters. The raters might have applied the consensus of scoring criteria they have obtained from the exercise to their scoring of S7-S12 for it happened right after the exercise. This could explain the reason why the inter-rater consistency of the first scoring of S7-S12 was high (.9) and also proved that the consensus building exercise did have an influence on teachers' inter-rater consistency to some extent. Even after one month, raters' inter-rater consistency was still more consistent (.81 for S1-S6, .88 for S7-S12). It appeared that the consensus building exercise has a lasting effect on teachers' inter-rater consistency.

By calculating the average of Pearson Product-Moment Correlation coefficient (γ) for each of the speaking components which were included in the grading criteria, including content, accuracy, fluency, pronunciation and vocabulary, teachers' inter-rater consistency for each component was also computed.

Table 4.6

Summary Table for Inter-Rater Consistency for Each Component and Total Score

	1st scoring of S1-S6	2nd scoring of S1-S6	1st scoring of S7-S12	2nd scoring of S7-S12
Content	.69	.66	.78	.83
Accuracy	.64	.66	.55	.62
Fluency	.64	.61	.73	.76
Pronunciation	.53	.55	.67	.71
Vocabulary	.61	.65	.63	.56
Total	.79	.81	.90	.88

In Table 4.6, the result indicates that inter-rater consistency for each component at different times was generally high. Correlations were slightly lower for accuracy in the first scoring of S7-S12. Lower correlations also happened in the first and the second scorings of S1-S6 for pronunciation and in the second scoring of S7-S12 for vocabulary.

Summarizing the part of the effect of the consensus building exercise on inter-rater consistency, the results reported that teachers' self consistency did increase after the exercise and the effect also lasted for one month. Such result is consistent with previous research showing that rater training could help raters to ease extreme point differences in scoring (Brown, 2003; Cook, 1989; Elder, Knoch, Barkhuizen, & Randow, 2005; Lumley & McNamara, 1995; Nakatsuhara, 2007). In addition, the result also disagreed with the conclusions made by other researchers (Congdon & McQueen, 2000; Weigle, 1994, 1998) indicating that rater training did not work well in increasing inter-rater consistency. As proposed by Congdon and McQueen (2000) and Elder et al. (2005), the problem of rater training's not working very well in improving inter-rater consistency can be solved by receiving peer feedback from group discussion and having ongoing training after a period of time. Consequently, the result is also consistent with what Congdon and McQueen (2000) and Elder et al. (2005) reported in their studies showing that raters' inter-rater consistency could still increased by using group discussion to get peer feedback and by holding a follow-up activity after one month.

The Effects of Short-term Consensus Building Exercise on Intra-rater Consistency

In this section, the researcher presents the results to address the second research question: *To what extent an individual rater was able to consistently assign the same scores to each video clip? In other words, how did the intra-rater consistency of the scores change after the consensus building exercise?*

It should be noted that, the first scorings for S1-S6 happened before the consensus building exercise. Immediately after the exercise, the raters began with their first scorings for S7-S12. The second scorings for S1-S6 took place one month after the exercise, so did the second scorings for S7-S12. In other words, raters' second scorings for both S1-S6 and S7-S12 happened on the same day; they all took place one month after the exercise.

By subtracting the mean score of the second scorings for a set of video clips from the mean score of the first scorings for the same video clips, the point differences gives an indication of intra-rater consistency. In other words, the lower point differences that a rater has between the first scorings and the second scorings, the higher the intra-rater consistency will be. For the content component, greater than 1 point difference was an extreme difference; for the other four components, greater than 0.5 point difference was an extreme difference; for the total score, greater than 3 points was considered as an extreme difference. Table 4.7 and Table 4.8 present the point differences for each component and total scores of S1-S6 and S7-S12 separately.

Table 4.7 shows the point differences for each component and for total scores of S1-S6. For new teachers, the point differences of rater N1's scores for content, accuracy, pronunciation and vocabulary was fairly small. Only in fluency component, the point difference of N1's scores was bigger and was considered as an extreme case. As for N2, the point differences of his scores for each component and for the total score were quite small, so N2 was pretty consistent throughout his first scorings and the second scorings. N3 only showed his biggest point difference in the content component, and the point difference of the total score was also considered as an extreme difference. N4 were quite inconsistent in every component, and in total score. On the contrary, N5 was consistent in every component, and in total score.

As for experienced teachers, they were generally more consistent than new raters. For example, E1 was only less consistent in vocabulary component; E4 was only less consistent in fluency component. E2 and E3 showed no extreme case of point differences in every component and total score; they were fairly consistent both the first time and second time. Only E5 showed two extreme point differences in accuracy and pronunciation.

In short, the comparison between the first scorings of S1-S6 (before the consensus building exercise) and the second scorings of S1-S6 (one month after the exercise) reveals the result that experienced teachers were generally more consistent than new teachers. In fact, two new teachers were deemed to have extreme total scores whereas none of the experienced teachers had extreme total scores. The result might also imply that there was a stronger influence of the consensus building exercise on new teachers.

Table 4.7

Point Differences for each Component and Total Scores of S1-S6

S1-S6	Content	Accuracy	Fluency	Pronunciation	Vocabulary	Total
N1	0.83	0	*1	0.17	0.17	2.17
N2	0	0	-0.33	0.50	-0.33	-0.17
N3	*-1.50	-0.50	-0.33	-0.33	-0.50	*-3.17
N4	*-1.50	*-0.83	*-0.83	*-0.83	*-0.67	*-4.67
N5	0	-0.50	-0.17	-0.50	-0.33	-1.50
E1	-0.25	0.50	0.50	0.17	*0.75	1.67
E2	0.17	-0.33	0.17	0	0	0.17
E3	1	0.50	0.50	0	0.17	2.17
E4	0.83	0.17	*0.67	-0.17	0	1.33
E5	0.17	*-0.58	-0.17	*-0.58	-0.42	-1.58

Note: An asterisk refers to an extreme point difference.

Table 4.8 shows the point differences for each component and for total scores of S7-S12. For new teachers, N1, N2, N3, and N4 showed no extreme point differences in each component and total score, which means that their scores for these five components and the total score were consistent both first time and second time. The first scorings of S7-S12 and the second scorings of S7-S12 all happened after the exercise, the only difference was that the second scorings of S7-S12 was hold one month after the exercise. Therefore, the comparison between the first scorings and the second scorings of S7-S12 gives an estimate of whether the exercise still affected these teachers after one month.

Table 4.8 reveals that only N5 showed the biggest point difference in accuracy, fluency, and pronunciation and in total score. As for experienced teachers, only E1 and E5 showed no extreme point difference. The point difference of E2's score for accuracy at the second time was 0.67 point deviated from the first time, which was considered as an extreme difference. E3 also showed a bigger point difference in accuracy as well. E4 showed bigger point differences in content, fluency and so did in total score. Compared with the result in Table 4.7, Table 4.8 shows that only one new teacher showed an extreme point difference in the total score whereas none of the experienced teachers showed an extreme point difference in the category of total score.

Table 4.8

Point Differences for each Component and Total Scores of S7-S12

S7-S12	Content	Accuracy	Fluency	Pronunciation	Vocabulary	Total
N1	0.50	-0.17	-0.17	0	-0.17	0
N2	0.33	-0.50	0	-0.33	-0.17	-0.67
N3	0	-0.33	-0.17	-0.33	-0.33	-1.17
N4	-0.50	0	-0.17	0.17	-0.25	-0.50
N5	-0.83	*-1.67	*-0.83	*-0.67	-0.50	*-4.50
E1	-0.08	-0.17	0.17	0.17	-0.08	0.17
E2	-0.33	*-0.67	0	-0.17	-0.50	-1.83
E3	0.17	*-1.17	-0.17	0.33	0	-0.83
E4	*1.17	0.50	*0.83	0	0.17	2.83
E5	0.17	0.25	0.25	0.25	0.33	1.17

Note: An asterisk refers to an extreme point difference.

In summary, the comparison between the first scorings of S7-S12 (immediately after the consensus building exercise) and the second scorings of S7-S12 (one month after the exercise) reveals that new teachers tended to improve their intra-rater consistency after the consensus building exercise, while the experienced teachers remained their consistency without being really influenced by the exercise. Especially rater N4 benefited the most after the exercise because the case of extreme point differences of N4's scores dropped from six to zero. Only one new rater showed an extreme point difference in total score after the exercise. As for experienced teachers, they showed fewer changes in their scorings after the exercise. The result might imply that new teachers changed more in their scores than the experienced teachers after the exercise. In other words, new teachers were influenced by the exercise more deeply than experienced teachers. Moreover, the comparison of the point differences between the first scorings and the second scorings of S7-S12 also discloses that the effect of the consensus building exercise did remain even one month later.

This finding is also consistent with the prior studies indicating that rater training appears to help raters to monitor their own consistency in order to reach higher intra-rater reliability (Alderson, Clapham & Wall, 2002; Brown, 2003; Lumley & McNamara, 1995; Weigle, 1994, 1998). The result of this study strengthened the effect of such short-term training workshop on increasing intra-rater consistency. Furthermore, the result indicated that the effect of improving inter-rater consistency through such a short-term training workshop could remain for one month. This result could be a new finding which the previous studies have not declared.

The Analysis of the Interview Data Collected from the Raters

In order to better understand the effects of short-term consensus building exercise, the researcher interviewed the ten raters individually after the experiment. The interviews began with the rater's background by asking the following questions: How many years have you taught in Tunghai University; Have you ever taught in any other school, if yes, how many years? Afterwards, some short questions were asked to trigger raters' response to the consensus building exercise. The interview questions are included in Appendix G.

The interview data not only reflected both new raters and experienced raters thoughts about the effects of the consensus building exercise, but also shed light for a future study of how to make such a kind of consensus building exercise better. The discussion of the interview data was divided into three parts: first, the effect of the consensus building exercise; second, the benefits that the raters gained from the consensus building exercise, and finally the raters' opinions or suggestions about the consensus building exercise.

The Effect of the Consensus Building Exercise

Most of the raters believed that they tried to keep their self consistency regardless whether they participated in the consensus building exercise; eight out of ten teachers especially mentioned that the grade they gave to the students were usually quite consistent regardless of the effect of the consensus building exercise. For example, N1 mentioned that “the scores I gave to my students were quite consistent regardless whether I attended the short-term consensus building exercise.” and E2 also indicated that “Well, I think I’m pretty consistent in my grading, for your consensus building exercise and in my class, too.”

But the raters still expressed that the consensus building exercise raised their concerns about the issue of being consistent even after a period of time. All of the raters mentioned the one thing that the consensus building exercise did help them to reach agreement among the other raters to some extent. They would try to revise their own scoring based on the consensus they came up with during the discussion in order to reach inter-rater consistency in the consensus building exercise. However, five out of ten raters also mentioned the fact that even though they did keep trying not to give too dissimilar scores from other raters in the consensus building exercise, but in their future classroom teaching, they would probably not be concerned about this problem. That is because each rater taught different level of students and their criteria of scoring or the activities in the classroom were basically different, so there was no way to see if the raters had a high inter-rater reliability or not (“…for example, a teacher who is teaching low level students will definitely not have the same criteria with the teacher who is teaching high or middle level students. We designed different oral tests and different classroom activities, so we surely had different criteria…”). These raters mentioned that this effectiveness might just remain at that time because they used the same criteria during the exercise, but in their later evaluation of students in the class, it might be difficult to achieve the inter-rater consistency. Therefore, one rater indicated that there is a need to come up with common criteria of speaking assessment according to each level. This might be a good way to provide teachers with an opportunity to examine whether their scoring differ from the other teachers and to help students have a better understanding about their general performance.

It was also interesting to find a fact that it was much easier for new teachers to change their scoring criteria after discussing with other raters. Certain new teachers mentioned that they adjusted their scoring criteria after the exercise (“I found my scores for students were different from the other teachers…I did modify my most different scores after discussing with other people”). On the contrary, it was more difficult for some experienced teachers to change their own standards even after discussing with other raters (“…we have our own judgments. And even if we discuss them, I don’t think they will change much”*)*.

The Benefits from the Consensus Building Exercise

The biggest benefit that every rater stated was from the discussion in the consensus building exercise They believed this consensus building exercise could help them to see how different raters' scorings were different from themselves ("…it helped to check my consistency and how I'm consistent with other people because different teachers look at different things…"), and helped them monitor their own scoring and also become more consistent with the other raters ("…the most interesting thing was comparing my consistency with other teachers' consistency and finding that there was a way to reach group consistency even though each teacher has different concept").

Most of the raters found that every rater tended to focus on different components while giving scores to the video clips. One rater might find that he tended to focus on grammar, while the other raters emphasized fluency or other elements. For example, one of the new teachers mentioned that "…but take myself for example, at I paid more attention to content and accuracy…", while the other new teacher indicated that "I don't really care about students' pronunciation. As long as their pronunciation is clear and articulate enough, I will give them a high score." One of the experienced teachers stated that "I'm more interested in if the students can understand the questions and respond appropriately. And so sometimes because of my many years' experience, I don't listen to grammar problem." In short, the result showed that every teacher has different standard and criteria of evaluating each component.

Some raters also mentioned that this exercise helped them to have a self examination about whether their criteria were fair enough to students. In addition, most of the raters modified their own scores to avoid a huge diversity from the other raters in order to reach a group consensus. One of the raters explained how the consensus building exercise helped them avoid giving extreme scores to the high level and the low level students ("…we can reach a consensus about students' oral proficiency with the other rates, so I will avoid giving high level student a very high score and giving low level student a very low score.")

Raters' opinions or Suggestions about the Consensus Building Exercise

The raters also provided some useful suggestions for the consensus building exercise. These suggestions will be helpful for any future workshop like this consensus building exercise. First of all, the raters all mentioned that there were too many video clips to watch and score. Most of them felt exhausted after scoring these twelve students' video clips in one short workshop. Besides, two out of ten raters also mentioned about the problem of the scoring criteria that they used in the consensus building exercise; they thought the categories of criteria were too many. They said that they were too busy in calculating the score and evaluating students' performance at the same time. However, the other eight raters indicated such criteria were very similar as the criteria they used in their classroom, so they did not have any problem of using it ("it was too exhausting for the teachers because it was noon time, and we have to watch so many video clips and scored them. After the consensus building exercise, we all felt very tired" or "the consensus building exercise took too much time scoring the video clips, it was…it was kind of exhausting, especially at the noon time, it would be better if the shorter consensus building exercise was shorter.")

To conclude the raters' opinions about the consensus building exercise, a future study should come out with a more time efficient workshop which may include less time watching video clips at one time but more time in discussing about how they scored the students. Additionally, when using analytical criteria like the researcher used in this study, the raters can be informed that they don't need to calculate the total scores by themselves; otherwise, the raters will be distracted from too many things at the same time.

Certain result of the interview data in this study is similar with what Weigle (1994) has done in her study indicating that the teachers tended to have more concerns for rater agreement and pay more attention to their self consistency after attending a training workshop. Raters' positive attitude toward the training workshop in this study is consistent with the result of Weigle's (1994) study. However, Weigle's (1994) interview data was collected immediately after the workshop, so those raters' memories were still fresh. On the contrary, the interview data of this study was collected six month after the follow-up activity so that raters kind of forgot what they did during the workshop. Therefore, some of the results collected from Weigle (1994) interview data could not be proved in this study, such as revision of raters' expectations of students' performance and task.

The interview data still provided further investigation of rater suggestions about the workshop, such as raters' tiredness after the whole exercise and their unfamiliarity of using certain rating criteria. The interview data further reported raters' concerns from classroom teachers' point of view, which was the difficulty of reaching inter-rater reliability in their future classroom assessment. It was because teachers who teach different level of student would definitely have different scoring criteria, so that it was more appropriate to come out with common criteria according to students' level difference. These above-mentioned findings offered thorough considerations for the future studies that the previous researches did not really provide.

CHAPTER FIVE

CONCLUSIONS

This chapter concludes the study by summarizing the study and its major findings. In addition, pedagogical implications for English teaching and learning are also discussed. Finally, the limitations of the study and suggestions for future research are included.

Summary of the Study

This study was conducted to investigate the effects of the consensus building exercise on raters. The purpose of the study were: 1) to investigate the effect of short-term training by analyzing intra-rater and inter-rater reliability, 2) to trigger raters' opinions about such kind of raters' consensus building exercise through the interview with raters, and 3) to encourage the use of short-term rater training consensus building exercise.

The data were collected through the consensus building exercise and its follow-up activity one month later and raters' interview after the follow-up activity finished. Before the exercise, the teachers scored the first six video clips (S1-S6).

After the exercise, they scored another six video clips (S7-S12). One month after the exercise, the teachers scored the same twelve video clips (S1-S12) again without having any exercise. The participants were ten teachers from the Freshman English for Non-majors (FENM) program in FLLD of Tunghai University. The participants were also divided into two groups according to their teaching years; five were more experienced teachers who have been teaching for more than five years in the FENM program, while the other five were newer teachers who have been teaching for less than five years in the FENM program. In order to collect more thorough data, the raters were interviewed to talk about their opinions and suggestions for the consensus building exercise six months after the follow-up activity.

Summary of Major Findings

The following paragraph summarized the major findings with the reference to the research questions:

First research question asked: *Was there any difference in the agreement among raters' scores of video clips before and after consensus building exercise? In short, what effects did short-term consensus building exercise have on inter-rater reliability?* By calculating the average of Pearson Product-Moment Correlations, the result can be used as an indication of inter-rater consistency. The higher the correlations were, the higher the inter-rater consistency would be. The result shows that the effectiveness of the consensus building exercise in terms of increasing inter-rater reliability could be proved because the inter-rater consistency became higher after receiving the consensus building exercise. Additionally, the effect still remained one month after the exercise.

The second research question asked: *To what extent an individual rater was able to consistently assign the same scores to each video clip? In other words, how did the intra-rater consistency of the scores change after the consensus building exercise?* The result reveals that new teachers tended to be influenced more by the exercise. New teachers' intra-rater consistency did improve after the exercise. Even after one month, the effect of the consensus building exercise still lasted. On the contrary, experienced teachers were generally very consistent within themselves before or even after the consensus building exercise; the experienced teachers did not change much after the exercise. The result implies that the consensus building exercise might have a stronger influence of affecting intra-rater consistency on new teachers than on experienced teachers.

The last research question asked: *What were the raters' opinions and suggestions about the consensus building exercise? Did the raters benefit from the exercise?* In order to answer this question, the researcher interviewed the teachers after finishing the follow-up activity. To summarize the interview, most of the raters believed that they always tried to keep their intra-rater consistency regardless in the consensus building exercise. For inter-rater reliability, some of the raters were not really concern about the problem of inter-rater reliability because they thought the level of each class and the classroom activities they applied would differ in their use of the criteria. It was also found that new teachers tended to show stronger willingness of adjusting their scores for the video clips after the exercise, while the experienced teachers did not really change much in their scores for the video clips. What benefited the raters the most was the group discussion from the consensus building exercise because all raters considered the discussion helped them not only to learn from the other raters through feedback from their peers, but also modified their scoring criteria in the future. When the raters achieved a group consensus, they might also apply this consensus to their future scoring. Finally, the raters suggested that the future consensus building exercise should be much shorter; otherwise, that will be too exhausting.

Pedagogical Implications

The major findings of this study have some pedagogical implications for language teachers. First of all, the results of the study provide evidence that doing this kind of short-term workshop for raters has its necessity. The findings imply that the implement of the consensus building exercise can increase teachers' inter-rater reliability to some extent. In addition, this exercise can also have a stronger influence on the newer teachers' scoring criteria and thus increase their intra-rater consistency. The results also prove that the effect of the exercise still remain even after one month.

Moreover, the researcher extracted one of the activities of rater training programs into the consensus building exercise: the group discussion that can not only help raters to achieve group consensus but also help them to be aware of their own self consistency. In addition, the consensus building exercise also included in other group activities, such as group scorings and group reports. The use of these group activities appears to be a good way to provide more chances for teachers to exchange their opinions and ideas about giving scores to students.

Another implication for language teachers is that the employment of consensus building can also encourage the use of speaking assessment. Sometimes, teachers may worry that speaking assessment may come along with lower rater reliability; besides, a large scale rater training program is always time-consuming. Therefore, the result of this study implies that of the effectiveness of such kind of consensus building exercise can encourage classroom teachers to test students' oral proficiency in order to achieve test validity and at the same time still can remain rater reliability to some degree. In short, raters' consensus building is doable and has its advantage because the consensus that raters achieved through the exercise can bring fairness and reliability of the test result.

Limitations of the Study and Suggestions for Future Study

This study confirms that effectiveness of consensus building exercise. However, there are still some limitations of the study. First, the raters did not receive examination to prove that they were already consistent or inconsistent within themselves before having the consensus building exercise; therefore, the result of the effectiveness of the consensus building exercise might be influenced by the degree of raters' original consistency. A suggestion for the future study is to have a former inspection of raters' consistency before they receive the treatment of the exercise or any kind of rater training.

Second, the numbers of the participants were small. In this study, only ten teachers participated in this consensus building exercise, so the statistical procedures were limited. Therefore, the difference before and after implementing the exercise might not be claimed significant. The suggestion for future study here is to collect more participants and tried to have a better arrangement of the teachers' time because they all have to work on their normal teaching job.

Third, the background of the raters should also be taken into consideration. The management of raters' gender, nationality, or their teaching years or the experiences of scoring or attending a similar workshop can also become a factor that influence the result. Therefore, the future studies can analyze these differences among raters to see if there is any possible effect of these factors on the results.

Finally, even though the raters mentioned that the two-hour consensus building exercise was too exhausting, such exercise might not be enough in terms of a rater training workshop. So, it could be a dilemma for future researcher. The suggestion for future study is to find a way to strike a balance between the time of the exercise and the raters' endurance. It could be workable if the future researcher shorten the time of exercise but increase the frequency of it. For example, the exercise could take about thirty minutes for each time but have raters attend such short-term workshop for more than one time.

REFERENCES

ACTFL (1999). The ACTFL Proficiency Guidelines: Speaking (revised 1999). Yonkers, NY: ACTFL.

Alderson, J. C., Clapham, C., & Wall, D. (2002). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Bachman, L. F. (2005). *Statistical Analysis for Language Assessment.* Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman, & A. D. Cohen (Eds.), Interfaces between Second Language Acquisition and Language Testing Research (pp. 112-140). Cambridge: Cambridge University Press.

Brown, G. & Yule, G. (2001). Teaching the Spoken Language: An approach based on the analysis of conversational English. Cambridge: Cambridge University Press.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing 20*, 1-25.

Brown, H. D. (2000). *Principles of Language Learning and Teaching*. New York: Pearson Education.

Brown, H. D. (2004). *Language Assessment: Principles and classroom practice.* New York: Pearson Education.

Canale, M. (1984). Considerations in the testing of reading and listening proficiency. *Foreign language Annals 17*, 349-357.

Chafe, W. (1985). Linguistic differences produced by differences between speech and writing. In D. R. Olsen, N. Torrance, & A. Hilyard (Eds.), *Literacy and Language Learning: The nature and consequences of reading and writing* (pp.229-255). Cambridge: Cambridge University Press.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Reaching of English 18*, 65-81.

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes 24*(3), 383-391.

Channell, J. (1994). *Vague Language*. Cambridge: Cambridge University Press.

Cohen, A. D. (1994). *Assessing Language Ability in the Classroom.* Boston: Heinle & Heinle.

Congdon P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement 37*(2), 163-178.

Cook, S. S. (1989). Improving the quality of student ratings of instruction: A look at two strategies. *Research in Higher Education 30*(1), 31-45.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Retrieved 2000, from http://www.coe.int/T/DG4/Linguistic/CADRE_EN.asp

Davis, A., Brown, A., Elder, C., Hill, K. Lumley, T. & McNamara, T. (1999). *Studies in Language Testing: Dictionary of language testing.* Cambridge: Cambridge University Press.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tests: *What does the test-taker have to offer? Language Testing 19*(4), 347-368.

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

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

ETS (2007). TSE and SPEAK Score User Guide. 2001-2002 edition. Princeton, NJ: Educational Testing Service. Online version available from http://www.toefl.org/tse/tseindex.html.

Genesee, F., & Upshur, J. (1999). *Classroom-based Evaluation in Second Language*

Education. Cambridge: Cambridge University Press.

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

Harris, D. P. (1986). Testing English as a Second Language. New York: MacGraw Hill Book Company.

Hasselgren, A. (1998). *Smallwords and Valid Testing*. Unpublished doctoral dissertation, University of Bergen, Bergen, Norway.

Henning, G. (1987). *A Guide to Language Testing.* Los Angeles: Newbury House.

Hughes, A. (2003). Testing for Language Teachers. Cambridge: Cambridge University Press.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F. & Hughey. J. B. (1981). *Testing ESL composition: A practical approach.* Rowley, Massachusetts: Newbury House.

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. New York: Prentice Hall.

Leather, J. & James, A. (1996). Second language speech. In C. William, & K. Bhatia (Eds.), Handbook of Second Language Acquisition (pp. 269-316). San Diego, CA: Academic Press.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professional 13*(4), 425-444.

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

Luoma, S. (2005). *Assessing Speaking.* Cambridge: Cambridge University Press.

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

Morley, J. (1991). The pronunciation component in teaching English to speakers of other languages. *TESOL Quarterly 25*, 481-520.

Mitchell, R., & Myles, M. (2004). Second language learning theories. London: Arnold.

McCarthy, M. & Carter, R. (1995). Spoken grammar: What is it and how can we teach it? *ELT Journal, 49*, 207-218.

Mohani, T. & Mohtar, T. (2005). A new dimension in assessing oral communication in a foreign language context. In J. A. Foley (Ed.), *Teachers' Perceptions towards Oral Assessment and Their Implications for Teaching* (pp. 205-223). Singapore: SEAMEO Regional Language Center.

Morley, J. (1991). The pronunciation component in teaching English to speakers of other languages. *TESOL Quarterly, 25*, 481-520.

Murphy, J. M. (1991). Oral communication in TESOL: Integrating speaking, listening, and pronunciation. *TESOL Quarterly 25*(1), 51-57.

Nakatsuhara, F. (2007). Inter-interviewer variation in oral interview tests. *ELT Journal*. Retrieved June 4, 2007, from *http://eltj.oxfordjournals.org/cgi/content/full/ccm044v1*

Nattinger, J. & DeCarrico, J. (1992). *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Nunan, D. (1993). Task-based syllabus design: selecting, grading and sequencing tasks. In G. Crookes, & S. Gass (Eds.), *Tasks in a Pedagogical Context: Integrating theory and practice* (pp. 55-68). Clevedon: Multilingual Matters.

O'Malley, J. M. & Pierce, L. V. (1996). *Authentic Assessment for English Language Learners.* United States of America: Addison-Wesley Publishing Company.

Pennington, M. C. & Richards, J. C. (1986). Pronunciation revisited. *TESOL Quarterly, 20*, 207-225.

Pawley, A. & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards, & R. W. Schmidt (Eds.), *Language and Communication* (pp. 191-227). London: Longman.

Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly, 17*, 219-240.

Sims, J. (2005). A new dimension in assessing oral communication in a foreign language context. In J. A. Foley (Ed.), *A New Dimension in the Teaching of Oral Communication* (pp. 240-253). Singapore: SEAMEO Regional Language Center.

Spear, M. (1996). The influence of halo effects upon teachers' assessments of written work. *Research in Education.* Retrieved November, 1996, from *http://findarticles.com/p/articles/mi_qa3765/is_199611/ai_n8756129*

Towell, R., Hawkins, R. & Bazerguin, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17,* 84-191.

Tuckman, B. W. (1988). Testing for Language Teachers. San Diego: Harcourt Brace Jovanovich.

Underhill, N. (1991). *Testing Spoken Language: A handbook of oral testing techniques.* Cambridge: Cambridge University Press.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing 11*(2), 197-223.

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

Wood, D., Bruner, J., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89-100.

APPENDIX A

ACTFL Proficiency Guidelines-Speaking (ACTFL, 1999)

SUPERIOR

Speakers at the Superior level are able to communicate in the language with accuracy and fluency in order to participate fully and effectively in conversations on a variety of topics in formal and informal settings from both concrete and abstract perspectives. They discuss their interests and special fields of competence, explain complex matters in detail, and provide lengthy and coherent narrations, all with ease, fluency, and accuracy. They explain their opinions on a number of topics of importance to them, such as social and political issues, and provide structured argument to support their opinions. They are able to construct and develop hypotheses to explore alternative possibilities. When appropriate, they use extended discourse without unnaturally lengthy hesitation to make their point, even when engaged in abstract elaborations. Such discourse, while coherent, may still be influenced by the Superior speakers own language patterns, rather than those of the target language.

Superior speakers command a variety of interactive and discourse strategies, such as turn-taking and separating main ideas from supporting information through the use of syntactic and lexical devices, as well as intonational features such as pitch, stress and tone. They demonstrate virtually no pattern of error in the use of basic structures. However, they may make sporadic errors, particularly in low-frequency structures and in some complex high-frequency structures more common to formal speech and writing. Such errors, if they do occur, do not distract the native interlocutor or interfere with communication.

ADVANCED HIGH

Speakers at the Advanced-High level perform all Advanced-level tasks with linguistic ease, confidence and competence. They are able to consistently explain in detail and narrate fully and accurately in all time frames. In addition, Advanced-High speakers handle the tasks pertaining to the Superior level but cannot sustain performance at that level across a variety of topics. They can provide a structured argument to support their opinions, and they may construct hypotheses, but patterns of error appear. They can discuss some topics abstractly, especially those relating to their particular interests and special fields of expertise, but in general, they are more comfortable discussing a variety of topics concretely.

Advanced-High speakers may demonstrate a well-developed ability to compensate for an imperfect grasp of some forms or for limitations in vocabulary by the confident use of communicative strategies, such as paraphrasing, circumlocution, and illustration. They use precise vocabulary and intonation to express meaning and often show great fluency and ease of speech. However, when called on to perform the complex tasks associated with the Superior level over a variety of topics, their language will at times break down or prove inadequate, or they may avoid the task altogether, for example, by resorting to simplification through the use of description or narration in place of argument or hypothesis.

ADVANCED MID

Speakers at the Advanced-Mid level are able to handle with ease and confidence a large number of communicative tasks. They participate actively in most informal and some formal exchanges on a variety of concrete topics relating to work, school, home, and leisure activities, as well as to events of current, public, and personal interest or individual relevance.

Advanced-Mid speakers demonstrate the ability to narrate and describe in all major time frames (past, present, and future) by providing a full account, with good control of aspect, as they adapt flexibly to the demands of the conversation. Narration and description tend to be combined and interwoven to relate relevant and supporting facts in connected, paragraph-length discourse.

Advanced-Mid speakers can handle successfully and with relative ease the linguistic challenges presented by a complication or unexpected turn of events that occurs within the context of a routine situation or communicative task with which they are otherwise familiar. Communicative strategies such as circumlocution or rephrasing are often employed for this purpose. The speech of Advanced-Mid speakers performing Advanced-level tasks is marked by substantial flow. Their vocabulary is fairly extensive although primarily generic in nature, except in the case of a particular area of specialization or interest. Dominant language discourse structures tend to recede, although discourse may still reflect the oral paragraph structure of their own language rather than that of the target language.

Advanced-Mid speakers contribute to conversations on a variety of familiar topics, dealt with concretely, with much accuracy, clarity and precision, and they convey their intended message without misrepresentation or confusion. They are readily understood by native speakers unaccustomed to dealing with non-natives. When called on to perform functions or handle topics associated with the Superior level, the quality and/or quantity of their speech will generally decline. Advanced-Mid speakers are often able to state an opinion or cite conditions; however, they lack the ability to consistently provide a structured argument in extended discourse. Advanced-Mid speakers may use a number of delaying strategies, resort to narration, description, explanation or anecdote, or simply attempt to avoid the linguistic demands of Superior-level tasks.

APPENDIX A (cont.)

ADVANCED LOW

Speakers at the Advanced-Low level are able to handle a variety of communicative tasks, although somewhat haltingly at times. They participate actively in most informal and a limited number of formal conversations on activities related to school, home, and leisure activities and, to a lesser degree, those related to events of work, current, public, and personal interest or individual relevance.

Advanced-Low speakers demonstrate the ability to narrate and describe in all major time frames (past, present and future) in paragraph length discourse, but control of aspect may be lacking at times. They can handle appropriately the linguistic challenges presented by a complication or unexpected turn of events that occurs within the context of a routine situation or communicative task with which they are otherwise familiar, thoug at times their discourse may be minimal for the level and strained. Communicative strategies such as rephrasing and circumlocution may be employed in such instances. In their narrations and descriptions, they combine and link sentences into connected discourse of paragraph length. When pressed for a fuller account, they tend to grope and rely on minimal discourse. Their utterances are typically not longer than a single paragraph. Structure of the dominant language is still evident in the use of false cognates, literal translations, or the oral paragraph structure of th speaker's own language rather than that of the target language.

While the language of Advanced-Low speakers may be marked by substantial, albeit irregular flow, it is typically somewhat strained and tentative, with noticeable self-correction and a certain >grammatical roughness.= The vocabulary of Advanced-Low speakers is primarily generic in nature.

Advanced-Low speakers contribute to the conversation with sufficient accuracy, clarity, and precision to convey their intended message without misrepresentation or confusion, and it can be understood by native speakers unaccustomed to dealing with non-natives, even though this may be achieved through repetition and restatement. When attempting to perform functions or handle topics associated with the Superior level, the linguisti quality and quantity of their speech will deteriorate significantly.

INTERMEDIATE HIGH

Intermediate-High speakers are able to converse with ease and confidence when dealing with most routine tasks and social situations of the Intermediate level. They are able to handle successfully many uncomplicated tasks and social situations requiring an exchange of basic informatio related to work, school, recreation, particular interests and areas of competence, though hesitation and errors may be evident.

Intermediate-High speakers handle the tasks pertaining to the Advanced level, but they are unable to sustain performance at that level over a variety of topics. With some consistency, speakers at the Intermediate High level narrate and describe in major time frames using connected discourse of paragraph length. However, their performance of these Advanced-level tasks will exhibit one or more features of breakdown, such as the failure to maintain the narration or description semantically or syntactically in the appropriate major time frame, the disintegration of connected discourse, th misuse of cohesive devises, a reduction in breadth and appropriateness of vocabulary, the failure to successfully circumlocute, or a significant amount of hesitation.

Intermediate-High speakers can generally be understood by native speakers unaccustomed to dealing with non-natives, although the dominant language is still evident (e.g. use of code-switching, false cognates, literal translations, etc.), and gaps in communication may occur.

INTERMEDIATE MID

Speakers at the Intermediate-Mid level are able to handle successfully a variety of uncomplicated communicative tasks in straightforward social situations. Conversation is generally limited to those predictable and concrete exchanges necessary for survival in the target culture; these include personal information covering self, family, home, daily activities, interests and personal preferences, as well as physical and social needs, such as food, shopping, travel and lodging.

Intermediate-Mid speakers tend to function reactively, for example, by responding to direct questions or requests for information. However, they ar capable of asking a variety of questions when necessary to obtain simple information to satisfy basic needs, such as directions, prices and services. When called on to perform functions or handle topics at the Advanced level, they provide some information but have difficulty linking ideas, manipulating time and aspect, and using communicative strategies, such as circumlocution.

Intermediate-Mid speakers are able to express personal meaning by creating with the language, in part by combining and recombining known elements and conversational input to make utterances of sentence length and some strings of sentences. Their speech may contain pauses, reformulations and self-corrections as they search for adequate vocabulary and appropriate language forms to express themselves. Because of inaccuracies in their vocabulary and/or pronunciation and/or grammar and/or syntax, misunderstandings can occur, but Intermediate-Mid speake are generally understood by sympathetic interlocutors accustomed to dealing with non-natives.

APPENDIX B

. The Test of Spoken English Rating Scale (ETS, 2007)

60 Communication almost always effective; task performed very competently

Functions performed clearly and effectively

Appropriate response to audience/situation

Coherent, with effective use of cohesive devices

Use of linguistic features almost always effective; communication not affected by minor errors.

50 Communication generally effective; task performed competently

Functions generally performed clearly and effectively

Generally appropriate response to audience/situation

Coherent, with some effective use of cohesive devices

Use of linguistic features generally effective; communication generally not affected by errors.

40 Communication somewhat effective; task performed somewhat competently

Functions performed somewhat clearly and effectively

Somewhat appropriate response to audience/situation

Somewhat coherent, with some use of cohesive devices

Use of linguistic features somewhat effective; communication sometimes affected by errors.

30 Communication generally not effective; task performed poorly

Functions generally performed unclearly and ineffectively

Generally inappropriate response to audience/situation

Generally incoherent, with little use of cohesive devices

Use of linguistic features generally poor; communication often impeded by major errors.

20 No effective communication; no evidence of ability to perform task

No evidence that functions were performed

No evidence of ability to respond appropriately to audience/situation

Incoherent, with no use of cohesive devices

Use of linguistic features poor; communication ineffective due to major errors.

APPENDIX C

TSE Descriptors of Overall Features (ETS, 2007)

60 Communication almost always effective; task performed very competently

Speaker volunteers information freely, with little or no effort, and may go beyond the task by using additional appropriate functions.

- Native-like repair strategies
- Sophisticated expressions
- Very strong content
- Almost no listener effort required

50 Communication generally effective; task performed competently

Speaker responds with effort; sometimes provides limited speech sample and sometimes runs out of time.

- Sometimes excessive, distracting, and ineffective repair strategies used to compensate for linguistic weaknesses (e.g., vocabulary and/or grammar)
- Adequate content
- Some listener effort required

40 Communication somewhat effective; task performed somewhat competently

Speaker responds with effort; sometimes provides limited speech sample and sometimes runs out of time.

- Sometimes excessive, distracting, and ineffective repair strategies used to compensate for linguistic weaknesses (e.g., vocabulary and/or grammar)

- Adequate content
- Some listener effort required

30 Communication generally not effective; task performed poorly

Speaker responds with much effort; provides limited speech sample and often runs out of time.

- Repair strategies excessive, very distracting, and ineffective
- Much listener effort required
- Difficult to tell if task is fully performed because of linguistic weaknesses, but function can be identified

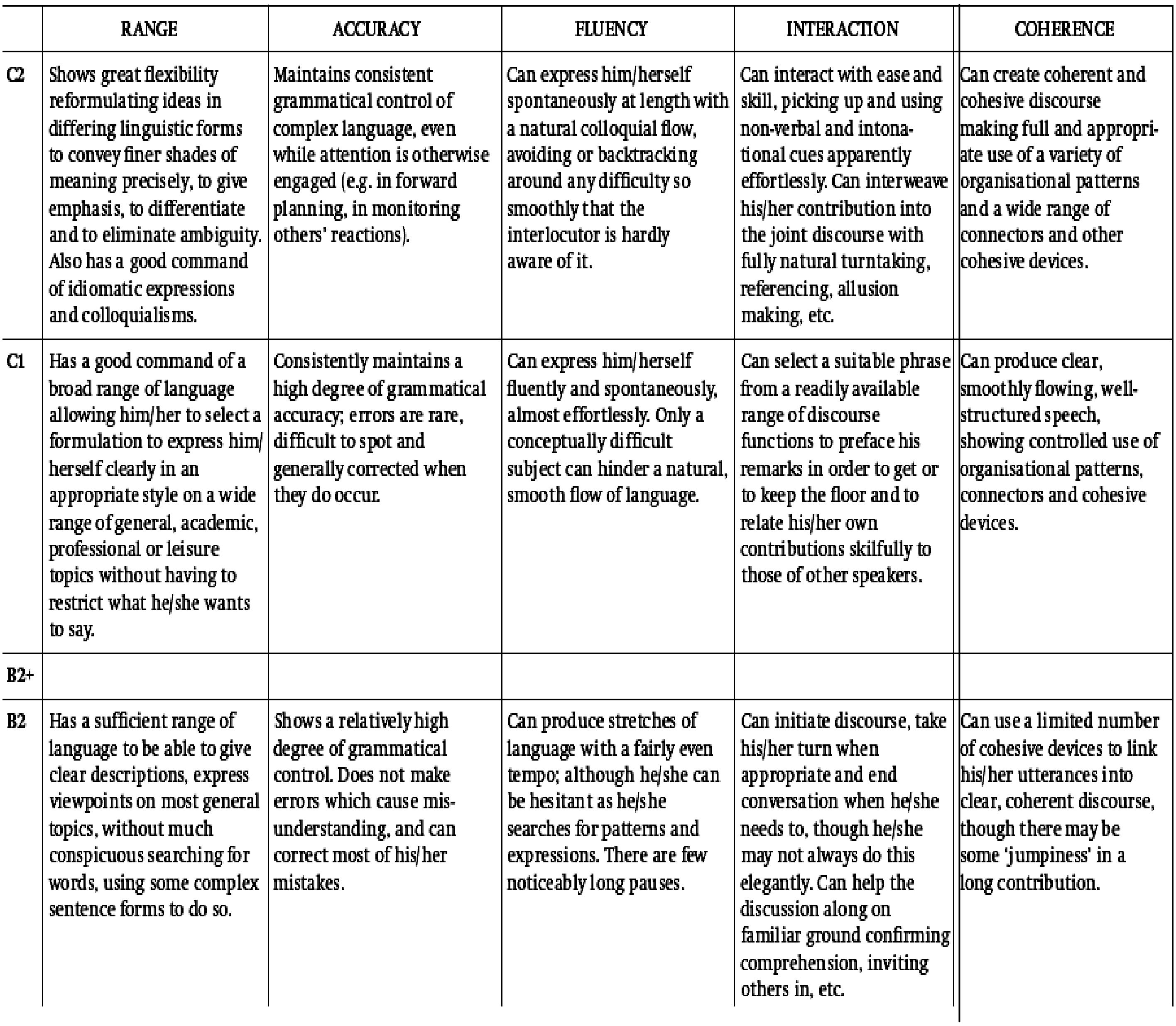
20 No effective communication; no evidence of ability to perform task

Extreme speaker effort is evident; speaker may repeat prompt, give up on task, or be silent.

- Attempts to perform task end in failure
- Only isolated words or phrases intelligible, even with much listener effort
- Function cannot be identified

APPENDIX D

Descriptors of Spoken Language (Council of Europe, 2001)



	RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turntaking, referencing, allusion making, etc.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skilfully to those of other speakers.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.
B2+					
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he/she searches for patterns and expressions. There are few noticeably long pauses.	Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some 'jumpiness' in a long contribution.

APPENDIX D (cont.)

B1+					
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.
A2+					
A2	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident.	Can answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.	Can link groups of words with simple connectors like 'and', 'but' and 'because'.
A1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorised repertoire.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.	Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.	Can link words or groups of words with very basic linear connectors like 'and' or 'then'.

APPENDIX E

Sample Grading Sheet

Students Name: ____________________ Student ID: __________________________

I. Content (10 points)

10 excellent, 9exceptional, 8 very good, 7 good,

6 acceptable, 5 poor, 4-1 failing ___________

II. Accuracy (5 points)

5 excellent, 4 good, 3 fair, 2 poor, 1 failing ___________

III. Fluency (5 points)

5 excellent, 4 good, 3 fair, 2 poor, 1 failing ___________

IV. Pronunciation ((5 points)

5 excellent, 4 good, 3 fair, 2 poor, 1 failing ___________

V. Vocabulary (5 points)

5 excellent, 4 good, 3 fair, 2 poor, 1 failing ___________

Total ___________

General Comments:

APPENDIX F

Oral Grading Guidelines

Content (10 points) – knowledgeable, substantive, thorough response, relevant. In short, a) does the student answer the question, b) does what the student says make sense, and c) is it relevant to the topic?

Accuracy (5 points) – tense, word order, complete sentences, pronouns

Fluency (5 points) – coherent and confident language use. This includes pronunciation, clarity, and speed of language. In short, can what the student says be understood?

Pronunciation (5 points) – clear and understandable. Can what the student says be understood?

Vocabulary (5 points) – appropriate register, effective word choice and usage, sophisticated range. In short, is the vocabulary appropriate and used correctly, and is there a range in vocabulary (does the student use more than just a few words)?

APPENDIX G

Interview Questions

Background:

How many years have you been teaching in Tunghai University?
How many years have you been teaching in Freshman English program?
Did you ever teach in any other school, if yes, how many years?

1. Did the short-term workshop work for you in maintaining your self consistency?

2. Do you think the short-term workshop generate agreement among raters to some extent?

3. Do you think the short-term workshop help you have more concern for inter-rater agreement and self consistency?

4. Did you benefit anything from the workshop?

5. Do you think the group discussion help you in any way?

6. Did you revise your expectations of examinee and task after the workshop?

7. Did you ever use any strategy or element that you've gained from the workshop in your grading afterwards?

8. What other elements do you think should be included in the grading criteria?

APPENDIX H

Raters' Scoring Results for S1-S6 from the First Time

S1

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	7	8	6	8	7	7	7	6	6	7	6.90
Accuracy	3	3	2	4	3	3	3	2	3	3	2.90
Fluency	3	3	3	4	3	2	3	2	2	3.5	2.85
Pronunciation	3	3	3	4	4	3	3	3	4	4	3.40
Vocabulary	3	3	3	3	3	3	4	2	3	4	3.10
Total	19	20	17	23	20	18	20	15	18	21.5	19.15

S2

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	6	8	7	7	7	8	8	6	6	6	6.90
Accuracy	3	4	4	4	4	4	4	3	3	3	3.60
Fluency	2	4	4	4	4	4	4	3	3	4	3.60
Pronunciation	2	4	4	4	4	4	3	3	4	4	3.60
Vocabulary	3	4	4	3	4	4	4	3	3	4	3.60
Total	16	24	23	22	23	24	23	18	19	21	21.30

S3

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	9	10	10	9	8	10	10	8	7	7	8.80
Accuracy	4	4	5	5	4	5	4	3	4	4	4.20
Fluency	4	4	4	5	4	4	4	3	4	4	4.00
Pronunciation	4	5	5	5	4	5	4	4	4	4	4.40
Vocabulary	4	4	4	5	4	4	4	4	4	4	4.10
Total	25	27	28	29	24	28	26	22	23	23	25.50

APPENDIX H (cont.)

S4

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	8	10	10	9	10	9	8	8	8	6.5	8.65
Accuracy	5	4	5	4	5	4	4	4	4	6	4.50
Fluency	5	5	5	5	5	4	3	4	4	4	4.40
Pronunciation	5	4	4	5	5	5	3	4	4	3	4.20
Vocabulary	4	5	5	4	4	4	4	4	4	3	4.10
Total	27	28	29	27	29	26	22	24	24	22.5	25.85

S5

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	6	8	8	8	7	8	7	5	6	6	6.90
Accuracy	3	2	3	4	3	3	3	2	3	3	2.90
Fluency	2	3	3	3	3	4	3	2	4	3	3.00
Pronunciation	3	2	3	3	3	4	3	3	3	3	3.00
Vocabulary	3	3	3	4	3	3	3	2	3	3	3.00
Total	17	18	20	22	19	22	19	14	19	18	18.80

S6

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	6	8	8	7	6	8	7	7	6	5.5	6.85
Accuracy	2	4	3	3	3	3	2	3	3	2	2.80
Fluency	2	3	3	3	3	3	2	2	3	2.5	2.65
Pronunciation	3	3	3	3	3	3	3	3	4	3	3.10
Vocabulary	3	4	3	3	3	3	3	2	3	2.5	2.95
Total	16	22	20	19	18	20	17	17	19	15.5	18.35

APPENDIX I

Raters' Scoring Results for S7-S12 from the First Time

S7

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	5	9	8	7	9	8	8	9	6	6	7.50
Accuracy	3	4	4	3	4	4	3	4	3	2.5	3.45
Fluency	4	3	4	3	5	4	3	5	3	3	3.70
Pronunciation	4	4	4	3	5	4	3	4	4	3	3.80
Vocabulary	3	4	3	4	5	4	4	4	4	3	3.80
Total	19	24	23	20	28	24	21	26	20	17.5	22.25

S8

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	5	7	6	5	6	7	7	5	5	5.5	5.85
Accuracy	2	3	3	2	3	3	2	5	2	2	2.70
Fluency	2	2	2	2	2	3	3	2	2	2	2.20
Pronunciation	3	3	3	3	2	3	3	2	3	2	2.70
Vocabulary	3	3	3	2	2	3	3	2	3	2.5	2.65
Total	15	18	17	14	15	19	18	16	15	14	16.10

S9

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	7	7	7	6	7	8	8	6	6	5.5	6.75
Accuracy	3	4	3	3	4	4	4	6	3	3	3.70
Fluency	2	3	3	2	4	4	4	3	3	3	3.10
Pronunciation	2	4	3	3	3	4	4	3	3	3	3.20
Vocabulary	3	4	4	3	3	4	3	2	3	3	3.20
Total	17	22	20	17	21	24	23	20	18	17.5	19.95

APPENDIX I (cont.)

S10

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	8	10	8	8	9	9	9	9	8	7	8.50
Accuracy	4	4	5	5	5	5	5	5	4	3	4.50
Fluency	5	4	5	5	5	4	4	5	4	3.5	4.45
Pronunciation	5	4	5	5	5	5	4	5	4	3	4.50
Vocabulary	5	5	5	5	4	5	5	5	4	3	4.60
Total	27	27	28	28	28	28	27	29	24	19.5	26.55

S11

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	9	9	9	8	10	9	8	10	8	7	8.70
Accuracy	4	4	5	5	5	5	4	5	4	3	4.40
Fluency	5	5	5	5	5	4	4	5	4	3	4.50
Pronunciation	5	4	5	5	5	5	4	5	4	3	4.50
Vocabulary	4	3	4	5	4	4	5	5	4	3	4.10
Total	27	25	28	28	29	27	25	30	24	19	26.20

S12

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	8	8	8	6	8	8	7	8	7	6	7.40
Accuracy	3	4	4	3	8	4	4	3	4	3	4.00
Fluency	4	4	4	3	3	4	3	3	3	3	3.40
Pronunciation	4	4	4	3	3	4	4	3	4	3	3.60
Vocabulary	3	4	4	3	3	4	4	3	4	3	3.50
Total	22	24	24	18	25	24	22	20	22	18	21.90

APPENDIX J

Raters' Scoring Results for S1-S6 from the Second Time

S1

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	7	8	6	6	7	6.5	9	7	7	7	7.05
Accuracy	3	3	3	3	3	3	3	3	3	3	3
Fluency	3	2	3	3	3	3	3	3	3	3	2.90
Pronunciation	3	4	3	3	3	3	3	3	3	2.5	3.05
Vocabulary	3	3	3	2	3	3.5	4	3	3	3	3.05
Total	19	20	18	17	19	19	22	19	19	18.5	19.10

S2

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	7	9	5	6	8	8	8	8	7	6	7.20
Accuracy	3	4	3	3	3	4	3	3	3	2.5	3.15
Fluency	3	4	3	3	4	3.5	4	3	3	3	3.35
Pronunciation	3	4	3	3	4	4	3	2	4	2	3.20
Vocabulary	4	4	3	3	3	4	4	3	3	3	3.40
Total	20	25	17	18	22	23.5	22	19	20	16.5	20.30

S3

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	9	10	9	8	8	9	9	9	8	7	8.60
Accuracy	4	4	4	4	3	5	4	4	4	4	4
Fluency	5	4	4	4	4	4.5	4	4	5	4	4.25
Pronunciation	4	5	4	4	4	5	4	4	4	4	4.20
Vocabulary	4	4	4	4	4	5	4	3	4	3	3.90
Total	26	27	25	24	23	28.5	25	24	25	22	25

APPENDIX J (cont.)

S4

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	9	9	8	8	9	9	8	9	8	7	8.40
Accuracy	4	4	4	4	4	5	4	4	4	3	4
Fluency	5	4	5	4	4	5	3	4	5	4	4.30
Pronunciation	5	4	4	4	3	5	4	4	4	3	4
Vocabulary	4	4	4	4	3	5	4	4	4	4	4
Total	27	25	25	24	23	29	23	25	25	21	24.70

S5

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	7	8	6	6	6	8	7	6	7	6	6.7o
Accuracy	3	3	3	3	3	4	2	3	3	2.5	2.95
Fluency	4	3	3	2	3	4	3	2	4	3	3.10
Pronunciation	3	4	3	3	3	4	3	3	3	3	3.20
Vocabulary	3	3	2	2	3	4	3	2	3	2.5	2.75
Total	20	21	17	16	18	24	18	16	20	17	18.70

S6

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	8	8	6	5	7	8	7	7	7	6	6.90
Accuracy	3	3	2	2	3	4	2	3	4	2.5	2.85
Fluency	4	3	2	3	3	4	3	3	4	3	3.20
Pronunciation	3	3	3	2	3	4	3	4	3	3	3.10
Vocabulary	3	3	3	3	3	4	3	3	3	2.5	3.05
Total	21	20	16	15	19	24	18	20	21	17	19.10

APPENDIX K

Raters' Scoring Results for S7-S12 from the Second Time

S7

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	8	9	8	7	8	9	7	9	8	6.5	7.95
Accuracy	3	3	4	4	3	4.5	3	4	4	3	3.55
Fluency	4	4	4	4	3	5	4	4	4	3.5	3.95
Pronunciation	4	3	4	4	4	4.5	4	5	4	3	3.95
Vocabulary	3	4	4	4	4	4.5	3	4	4	3.5	3.8
Total	22	23	24	23	22	27.5	21	26	24	19.5	23.2

S8

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	6	7	5	5	6	7	6	6	7	6	6.10
Accuracy	2	2	2	3	3	3	2	2	4	2.5	2.55
Fluency	2	2	2	2	2	3	3	2	3	2.5	2.35
Pronunciation	3	3	3	3	2	4	3	3	3	2.5	2.95
Vocabulary	3	3	2	2	2	3	2	2	4	3	2.60
Total	16	17	14	15	15	20	16	15	21	16.5	16.60

S9

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	6	8	6	5	7	7.5	7	7	7	5.5	6.60
Accuracy	2	3	3	3	3	3.5	2	3	3	2.5	2.80
Fluency	2	3	3	2	3	3.5	3	3	4	2.5	2.90
Pronunciation	3	3	3	3	3	3.5	3	3	3	3	3.05
Vocabulary	3	3	2	2	3	3.5	3	3	3	2.5	2.80
Total	16	20	17	15	19	21.5	18	19	20	16	18.20

APPENDIX K (cont.)

S10

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	9	9	10	8	8	9	9	9	9	7	8.70
Accuracy	5	4	4	4	4	5	4	5	4	3.5	4.25
Fluency	5	4	5	4	4	4	4	5	5	3	4.30
Pronunciation	5	4	4	5	3	5	4	5	4	3.5	4.25
Vocabulary	4	4	5	4	3	5	4	5	4	3	4.10
Total	28	25	28	25	22	28	25	29	26	20	25.60

S11

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	8	10	9	8	8	9	8	9	9	7	8.50
Accuracy	3	4	5	4	3	5	4	4	4	3	3.90
Fluency	4	4	4	4	4	4.5	4	5	4	4	4.15
Pronunciation	4	4	4	5	4	5	4	4	4	3.5	4.15
Vocabulary	3	4	4	5	3	5	4	4	5	4	4.10
Total	22	26	26	26	22	28.5	24	26	26	21.5	24.80

S12

Rater	N1	N2	N3	N4	N5	E1	E2	E3	E4	E5	Mean
Content	8	9	8	6	7	7	8	8	7	6	7.40
Accuracy	3	4	4	3	3	3	3	3	4	3.5	3.35
Fluency	4	4	4	3	3	4	3	3	4	3.5	3.55
Pronunciation	4	4	4	3	3	4	3	4	4	3	3.60
Vocabulary	4	4	4	3	3	3.5	4	3	4	3	3.55
Total	23	25	24	18	19	21.5	21	21	23	19	21.50