INDIVIDUAL AND PAIRED ORAL PROFICIENCY TESTING: A STUDY OF LEARNERS' PREFERENCE

By

Koffi Nicolas Kanga

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF ARTS

Teaching English to Speakers of Other Languages

2012

ABSTRACT

INDIVIDUAL AND PAIRED ORAL PROFICIENCY TESTING:
A STUDY OF LEARNERS' PREFERENCE

By

Koffi Nicolas Kanga

Since the introduction of paired testing, oral proficiency tests for English as a Second Language speakers (ESL) has consisted of an individual test, a paired test, or a combination of both formats depending on institutions and their stakeholders. The present study aims at investigating which of both exam formats English as Second Language speakers and English as a Foreign Language (ESL/EFL) speakers prefer, and the reasons behind the preferences. Data were collected from 15 ESL students at a large university in the midwestern United States. Participants took a paired and an individual test; following the tests participants were interviewed on their preferences and their reasons for these preferences. Results showed that first the majority of participants preferred the individual test, secondly that the participants' reasons behind their choices related to the ease of the test and their state of emotional comfort during the test. However, while participants mostly preferred the individual test, their test results show that they did not perform better in that test format. The results have implications for the preparation of ESL students and future candidates for oral proficiency tests, but suggest that more studies should be conducted for a better understanding of learners' preferences and its implications for the assessment of oral proficiency.

.

DEDICATION

Dedicated to Irene Kanga and my children

with love from Nicolas

ACKNOWLEDGEMENTS

Writing a thesis could be compared to climbing a mountain. Where there are easy grips motivation is high and progress fast. But many times temptation to give up is big; and it takes people with faith and knowledge to help you continue. Many people played this role for me on this thesis, they helped me through the process and I would like to take this opportunity to thank them.

First and foremost, I wish to thank my advisor, Dr. Debra Friedman, whose dedication and patience never cease to amaze me. With her expertise and insightful comments, she has patiently guided me from the beginning. In addition to providing support throughout my thesis, she has guided my first steps into the academic community; she has given me more than I can confess. I express my sincerest thanks and appreciation.

Next, I would like to thank Dr Paula Winke, my second reader, for her advice and encouragement. It was in one of her classes that I picked the idea for this thesis and she guided the first steps. I'm very grateful Paula.

I would also take advantage of this opportunity to express my appreciation to the faculty of the MA-TESOL program at Michigan State University. Dr Debra Hardison, Dr Charlene Polio, Dr Shawn Loewen, Dr Patti Spinner it was an honor and privilege for me to be your student. Sometimes I looked slow, it was because I wanted to get the most of your instruction.

My gratitude also goes to my fellow Moise Konate. He was a support throughout this work. Given the design of this thesis I certainly could not continue if I didn't have him. Thank you Moise for being the second researcher is my research. May God bless you.

My appreciation also goes the administration, the teachers, and the students of the English Language Center (ELC) at Michigan State University, where I collected my data. I would like to thank particularly Anne Marie Desiderio for her support and contribution.

This acknowledgement would not be complete without expressing my gratitude to the Fulbright exchange program. This thesis would not have been possible had it not been for the invaluable support of the officials in charge of the Fulbright exchange program. I would especially like to express my deepest gratitude to the U.S Department of State for the opportunity I was given to discover The American educational system and to graduate from such a prestigious university as Michigan State University.

Finally I would like to thank my family and all my friends from afar and those geographically closer, who provided advice support and encouragement. To all these people I would like to say thank you and make a promise to put their advice and support to good use.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 – Introduction and Literature Review

Introduction

Although the testing of speaking in second/foreign language teaching has a long history it was not until the 1980s that it became commonplace (Spolsky, 2001, cited in Alderson & Banerja, 2002, p. 92). The spread of speaking tests occurred with the advent of communicative testing, which no longer limited language skills to reading and grammar knowledge (Richards & Rodgers, 2001), and the growing importance of the English language around the world (Hsu, 2009).With the growing interest for oral proficiency testing then, different formats of oral proficiency tests were developed and tried. Those new oral tests included the Oral Proficiency Interview (OPI) developed by the Foreign Service Institute (FSI) and Associated Government Agencies, the (ACTFL-OPI) developed by the American Council on the Teaching of Foreign Languages (ACTFL), and the paired oral test format developed by Cambridge.

This period of the 1980s corresponded to an era of intensive research on oral proficiency tests. This research was concerned with the capacity of oral tests to allow accurate judgment of learners' proficiency, how to best assess speaking proficiency, the best ways to elicit speaking samples from learners, and the effect of interlocutors on scores and performances. It is this research which lead researchers associated with Cambridge English as a Second Oral Language (ESOL) to introduce the paired format of oral proficiency testing. The two main concerns underlying to the adoption of the paired oral test format were elicitation and rating; researchers were concerned with the best ways to elicit spoken language from learners and how to rate learners' performances accurately. According to Saville and Hargreaves (1999) the paired format of the oral proficiency test allows more variety of patterns of interactions (examinee – examiner

and examinee – examinee) and more fairness in scores, since candidates are judged in two different ways by two examiners. (Saville & Hargreaves, 1999, pp. 43-44)

Oral proficiency tests in second language learning contexts have often played the role of gatekeeping tests as people move around the world in search of jobs and opportunities. The purpose of such tests has generally been to assess whether the test taker's linguistic proficiency is adequate for a predetermined assignment (Jenkins & Parra, 2003, p. 90). As such, oral proficiency tests play an important role in the lives of many second language (L2) speakers; they may allow or upset their chances to register for further studies in English speaking universities, their plans to work for international organizations, or for multinational companies (Hsu, 2009).

Many oral proficiency examinations follow the ACTFL-OPI format; that is, the party whose 'spoken language' is the object of the assessment is a nonnative speaker of the second or foreign language, which is both the (primary) medium of the interaction and the assessment target. The candidate interacts with a certified assessor, an expert speaker of the target language who conducts the interaction in an interview-type fashion according to a pre-determined, but unscripted, protocol. The charge of the interviewer is to elicit speech samples from the candidate that enable ratings of the candidate's performance on a pre-specified scale, such as the ACTFL Proficiency Guidelines (American Council on the Teaching of Foreign Languages, 1999).

The paired format proposed by Cambridge ESOL adopts a different approach. Rather than having the test-taker interact with an expert speaker, the test-taker interacts mainly with a peer in front of two examiners. One of the two examiners the *interlocutor* occasionally interacts with candidates, mainly to set the tasks, while the second examiner referred to as the *assessor* remains silent all the time. The main task of the *assessor* is to listen to the candidates and assess

them on the evidence of their performance in the tasks, against the established criteria (Foot, 1999, p. 39).

The paired oral test format has won the favor of many institutions. In addition to Cambridge ESOL, which has introduced the paired test format in all its examination suites, the paired test has been adopted by many universities and government examination boards. Csepes (2005) and Együd and Glover (2001) provide an example of the introduction of paired testing in Hungarian matriculation examinations at secondary school level. Another example of the adoption of paired testing is provided by Brooks (2009), who reports the decision of the administration of an English for Academic Purposes (EAP) institute in Canada to introduce paired testing in their end of the training exam suite.

Due to the importance of oral proficiency tests, the techniques and methods used to deliver those tests deserve attention; all factors and variables involved in the tests need to be carefully investigated in order to determine "whether or not these variables, through their effects on discourse, impact the scores assigned in these situations" (Lazaraton, 2006, p. 287). It is certainly for this reason that many research studies have been conducted and published on both test formats. Unfortunately, learner preference has received very little attention in research, although many researchers have advocated research studies on this topic since the introduction of paired testing (Foot, 1999; Norton, 2005). Research in this area should contribute knowledge about learners' attitudes towards those tests, data for improving oral tests, and ways to treat candidates fairly.

The present study explores learners' preference between the individual test format used by the International English Language Testing System (IELTS) and many oral proficiency examinations and the paired test format introduced by Cambridge Local Examinations Syndicate (UCLES) and used in the First Certificate of English (FCE) and all their exam suites. It is inspired by the limited number of publications on this issue and the apparent contradictions in existing results on learner preference (Együd & Glover, 2001; Marochi, 2008; Taylor 2001). Through a qualitative approach the study investigates the choices made by participants after taking the two types of tests. The main concern of the study is whether candidates have a preference for one of the two test formats, the reasons behind their choices, and the implications of these preferences.

Statement of the problem.

Far from settling debates about oral proficiency tests, the paired test format seems to have introduced new ones. Saville and Hargreave (1999) claimed that the aim behind the introduction of the paired test was to be fair to candidates. Yet many researchers are doubtful whether this objective has really been achieved; indeed, the paired test format has introduced in the oral testing process new factors, such as a second speaking partner, a second scorer, and a new type of task, whose impact on candidates' production is widely admitted (Brooks, 2009; Davis, 2009; Lazaraton, 2006). For example Brooks (2009) and Davis (2009), show that when working in pairs learners produce more language, which results in better grades. However, no systematic data analysis has shown the magnitude of the impact of those factors on scores awarded. This

fact raises issues about the validity and reliability of the new test and whether the new test is really fair to candidates. Researchers (Foot, 1999; Norton, 2005) conclude that "At the moment it looks as though in addressing one problem, the examiners have created several new ones"(Foot, p. 41). Unfortunately no data on preliminary implementations conducted by UCLES Prior to the introduction of the paired test format are available for researchers.

Secondly, the existence of two test formats in oral proficiency testing raises the question of test-taker preferences. Which test format candidates would like to take? As stated earlier, the test population for oral proficiency tests consists essentially of nationals from countries where English as a Second Language /English as a Foreign Language (ESL/EFL) teaching and testing is still dominated by traditional methods (Sawir, 2005). Sawir writes that "traditional EFL pedagogies in East and Southeast Asian nations… focus almost exclusively on learning to read English-language documents, and to prepare English language essays and letters, with little attention to the skills of conversation in English, let alone the ultimate communicative goal of native speaker-level proficiency". (p. 567-568). So it is not wrong to assume that for such a population which is not used to producing English orally, spoken tests in general is a potential cause of worry and anxiety. Such a population will certainly want to choose the test type in which it will feel more comfortable.

Finally, the issue of preferences is an important one because of the potential impact of candidates' preferences and the factors underlying those preferences on performances and grades. According to literature, factors like test difficulty and face validity have significant impact on grades and performances. Hong (1999) notes that: "the perceived difficulty of the exam by students arouses students' worry, which in turn affects their performance" (p. 432). Equally, much other work in psychology reports similar relationship between perceived task

difficulty and students' low performance owing to the feeling of anxiety caused by this perception. (Fairclough, Tattersall & Houston, 2006; Horwitz, 1995; Kivimäki, 1995).

Learners' performances may also be affected by their perception of the validity of the test. Test-takers' perception of the test is referred to as *face validity*, it is otherwise referred to as *test appeal* or *test appearance* (Bachman, 1990, p. 287). Defining the concept, Anastasi (1988) writes that "face validity is not validity in the technical sense: it refers not to what the test actually measures, but to what it appears superficially to measure. Face validity pertains to whether the test 'looks valid' to the examinees who take it, the administrative personnel who decide on its use and other technically untrained observers" (Anastasi, 1988, p. 144, cited in Kajdaz, 2010, p. 6).

Because it is considered as a subjective view about the test, a "layman's measure", face validity was bitterly criticized by specialists; many of those specialists cited by Bachman (1990) even suggested abandoning the concept of face validity because it is said to have no theoretical basis or scientific evidence. They contend that test validation should be left to experts with relevant training in test development and analysis. However, although they argued against it, many of them at the same time recognized that test appearance has a considerable effect on the acceptability of tests to both test takers and test users (Bachman, 1990). Indeed, test designers cannot ignore the opinion of test takers and test users, that is the concept of face validity, because they will always have to ensure that the test is acceptable to its users, and therefore that its results are taken seriously by all concerned (Elder, Iwashita, & McNamara, 2002). Further, beyond the idea of test acceptability, many researchers have reported the impact of face validity on candidates' performances. Test-takers' opinion about the validity of the test not only determines how seriously they will work during the test, but there is also a clear relationship between test-

takers' opinion about the validity of the test and performance. Bachman (1990) highlights this connection when he concludes that "the 'bottom line' in any language testing situation, in a very practical sense, is whether test takers will take the test seriously enough to try their best, and whether test users will accept the test and find it useful. For these reasons, test appearance is a very important consideration in test use" (Bachman, 1990, p. 288). Similarly, Brown (2007) writes that "to achieve peak performance on a test a learner needs to be convinced that the test is indeed testing what it claims to test" (p. 449).

In brief, because of all that is involved in oral proficiency tests, such as learners' education background and learners' perception of test validity and difficulty, it seems questionable for test experts and test users to impose one test type without taking into account learners' preferences and their reasons. The study of learners' preferences will provide experts, with useful information for an informed decision.

Rationale of the study.

Research conducted in view of finding the best way to elicit spoken language from learners has found paired test to be a sound alternative, and this test format has won the favor of many universities and testing institutions. But how students feel about changing from the traditional individual test to the new testing approach has remained a mystery. Similarly little is known about which of the two test formats learners prefer. For institutions that are still hesitating between the two test formats, a study about learner preference will provide useful information to guide their choice.

Unfortunately, up to now there has been a contradiction between the data provided by previous research on learner preference. While one group of research has concluded that learners

prefer the paired test, results published by other studies have revealed that a preference for the individual test. Faced with the limited number of works on this topic and the contradictions in existing research results, the purpose of the present study is an attempt to provide a clear answer regarding the type of oral proficiency test participants prefer.

The study will also fill a gap. Faced with the lack of data from the implementations that were carried out prior to the generalization of paired testing, some researchers suggested the study of learner preference as a study orientation for collecting information. Foot (1999) suggested that such studies would generate data that may be useful for improving oral proficiency testing in general and individual testing in particular. By collecting information on learners' perceptions and attitudes the study expects to contribute along this line. The logical follow up to the study of learner preference, that is, the study of the reasons behind learners' choices will certainly serve this purpose.

Another reason for studying learners' preferences derives from the impact of those preferences on performances. Much research has concluded that learners' performances may be affected by their opinions about the test (Bachman, 1990; Brown, 2007). It is believed that learners will work better in the test format that they prefer, since they might show more motivation in that test. Therefore the study will investigate the relationship between the preferences expressed by participants and their actual performances on the tests.

Many questions regarding learners' preferences, their reasons for these preferences, and their implications have remained unanswered. Answering these questions may be helpful for learners, testing institutions, and for research on the testing of spoken language.

Research aims and objectives.

The study has three main objectives: actualize general knowledge on learners' preferences between the paired test and the individual test and to contribute to the database of arguments that may be taken into account to improve oral tests and testing conditions, in order to allow students to produce their best English. The study gives participants the opportunity to discover and experiment with paired testing and individual testing and to state their opinions about the type of test they would rather take if they were given the choice. Next the study investigates the reasons behind the choices made by learners. This involves getting participants to verbalize some of their feelings, their likes and dislikes during and after test sessions. Finally the study analyzes the relationship between the preferences expressed by each participant and his or her performances in the two tests.

Significance of the study.

Several aspects of oral testing have been under investigation in recent years in order to make sure that candidates' performances are accurately assessed, that candidates take the tests in appropriate conditions and perform at their best. It is the aim of the present study to contribute elements at some of these levels.

Lazaraton (2006) has acknowledged the fact that "variables such as gender, proficiency, and acquaintanceship affect the discourse produced in paired speaking tests"; yet she wonders whether they impact candidates' scores. One factor which could logically appear in the list suggested by Lazaraton is learner preference. Unfortunately much less has been written on this factor; the stress has so far rests on external factors. Therefore the question still remains to find

out whether learner preference has some impact on their productions and whether it would influence scores.

After more than two decades of existence of the paired test has appeared as a major trend in oral testing. It has therefore fueled many research studies on its performances and also comparative studies with the individual test. Some of these research studies have focused on learner preference. Although this body of research has provided arguments on the reasons underlying candidates' preferences it is the purpose of the present study to collect new or complementary arguments and factors which are deemed significant enough to influence the choices made by candidates between the individual and the paired test formats.

The study is also significant for ESL/EFL pedagogy; the coexistence of both test formats does not induce learners alone to make a choice. More and more, language institutions, universities and testing institutions are facing the need to take stances. Some are simply dropping the individual test to adopt the paired test as their sole test format while some are defending the values of the individual test format and keep it as their exclusive test scheme. In such a landscape it is of interest for research to provide elements for informed choice to teaching and testing institutions. Foreign language teachers may learn facts from learners' perspective that could help them maximize the chance of their learners to pass their oral proficiency tests.

Literature Review

Paired and individual testing: advantages and disadvantages.

Many papers have discussed the advantages and disadvantages of different oral tests, but whether one test format has more advantages than the other remains a controversial issue. To justify the generalization of paired testing to all levels of Cambridge exams, Saville and Hargreave (1999) have argued that on paired testing candidates are more relaxed, they have the possibility of more varied patterns of interaction during the tests, and this format can lead to positive washback in the classroom by encouraging learners to interact together in preparation for the test. These arguments are shared by Együd and Glover (2001), who believe that " it would be wrong to choose examination formats that reflect the unrealistic interaction patterns common to teacher-centered classrooms, since this would work against the sort of teaching that uses a wider range of interaction in class" (p. 75). Taking a more pragmatic perspective, Ducasse and Brown (2009) write that "peer-to-peer assessment is typically also more time and cost efficient as candidates are tested together, and raters assess two or more candidates simultaneously" (p. 424). For Brooks (2009) "paired testing more closely mirrors the type of oral interaction the students would likely encounter at university and it reflects the type of speaking tasks commonly used in the classroom" (p. 324). A similar argument is proposed by van Lier (1989), who believes a test format that is closer to conversation is better than the interview test because the interview tests resulted in 'test discourse' or 'institutional talk', and did not represent normal conversation.

For Foot (1999), however, even if there are some truths about the advantages mentioned, they do not justify the shift to paired testing; he thinks that this shift has complicated further the situation of oral assessment. His article is a list of potential risks and disadvantages of paired

testing. Dealing with candidates' anxiety, he argues that candidates may not necessarily perform 'better' if they are more relaxed; nervous candidates could make their partner feel more nervous; candidates who do not know each other may feel more anxious about interacting with a stranger. In other words, he does not believe that paired testing reduces candidate anxiety nor does he believe that candidates would perform better if they were not nervous. In addition, Foot claims that there is still not enough evidence about the effect of differences in language ability, L1 background, social relationships, and factors such as age, personality, and social class upon linguistic performance in the paired interview format. Norton (2005) agrees with Foot, mostly on the risks regarding pairing candidates with different linguistic abilities and candidates who do not know each other. Although she does not reject the idea of paired testing, she recommends caution; she believes many factors should be researched before paired testing is generalized. She concludes that "it would seem worthwhile to find out student preferences for paired or individual interviews in speaking tests. It would also be interesting to investigate examiner perceptions of the advantages and disadvantages of the paired format" (p. 295).

Candidate performance on paired and individual testing.

Comparative studies of candidate production on paired testing and individual testing reveal that learners perform better on a paired test than on an individual test. They get better scores as a result of a more complex grammar and vocabulary, a higher amount of talk produced and finally a better use of interaction devices. Marochi (2008) analyzed the amount of language produced, the complexity of grammatical structures used, and the diversity of vocabulary. She found that learners performed better on paired testing than they did on individual tests. Her conclusions are supported by other studies which show that when working in pairs learners produce more language, which results in better grades (Brooks, 2009; Davis, 2009). Brooks

compared the results of 16 candidates who took a paired test and an individual test and found that 12 out of the 16 participants got better marks on the paired test. Brooks' study, like the one conducted by Davis, reveals that learners engaged in paired testing display a greater variety of interaction patterns than they do when they interact with the examiner on the individual test. This conclusion is supported by Galaczi (2008); she noticed three kinds of interaction patterns that candidates use exclusively or together in different proportions when performing on pair testing. These patterns are collaborative, parallel, and asymmetric. Her results show that candidates who used the collaborative mode scored higher than those who used the other modes.

Interlocutor and partner effect.

A major issue affecting all face-to-face oral assessment is the presence and involvement of a second person or group of people with the candidate. This presence induces factors such as the partner's or the interviewer's age, gender, L1, and degree of acquaintanceship, which are believed to induce variance in scores obtained by candidates. Studies however, report controversial and sometimes conflicting results. Among other topics Porter (1991) studied the effect of gender in spoken tests. His implementations conducted at Reading University, UK with a group of Arabs and Algerians reveal that those students scored higher when they were interviewed by a male interviewer, and they performed at their best when that male interviewer was not marked with high social status. However, a study with similar design conducted by Porter and Shen (1991, cited in O'Sullivan, 2000, pp. 374-375) with Japanese students gave a different result; participants scored higher when they were interviewed by a woman. This last result is similar to O'Sullivan's (2000) conclusions; his study conducted with 12 Japanese students (six male and six female) revealed that candidates in general (irrespective of their gender) performed better with female interviewers. However, O'Loughlin (2000) and Brown and

McNamara (2004) did not find any significant evidence of the effect of gender on candidates' performances, even though Brown and McNamara found that male interviewers have a different interview style as compared to their female colleagues.

Another factor that has caught the attention of researchers is familiarity with the partner in paired tests or the interviewer in individual tests. Do candidates perform better when they are paired with a friend or when they are interviewed by someone they know? The study of Porter (1991) does not show any significant relationship. Yet, work in psychology show that "spontaneous support offered by a friend positively affects anxiety and task performance under experimental conditions" (O'Sullivan, 2002, p. 279). O'Sullivan's study of this issue produced three results. First, it shows that, contrary to the interview situation, in paired testing the partners' gender does not have any significant influence on candidates' scores. Secondly, the study shows that one can notice a difference in grammatical accuracy and complexity when subjects are paired with partners of different gender. Finally, contrary to Porter (1991), who found no such evidence, O'Sullivan reports that candidates' performance is affected by their degree of familiarity with their partners. Obviously, concerning the effect of acquaintance, there is a clear divide between the individual interview situation and the paired testing situation. The different studies do not reveal any significant impact of interviewer familiarity in individual testing. But in the case of paired testing interacting with a friend partner has been shown to have noticeable impact on scores.

The difference in speaking partners' ability is of particular concern in paired tests. Many researchers believe that the proficiency level of the partner in paired testing could be a potential source of score variance (Foot, 1999; Norton, 2005). Iwashita (1996) confirmed this hypothesis when she found that the interlocutor's proficiency affects both the amount of talk produced and

the scores. But other researchers reached different conclusions. Davis (2009) reports the study of Nakatsuhara (2004), who found that "in general, differences in proficiency level among candidates had little effect on conversation type, although in mixed parings of higher- and lower-proficiency individuals, the higher proficiency test taker spoke more and initiated more topics" (p. 370). Davis shares Nakatsuhara's conclusion; Davis's raw data tend to suggest that the partner's level could affect production, in terms of interaction pattern and amount of talk produced. But overall his work did not show any significant evidence; he concluded that "interlocutor proficiency may influence scores for some individuals, but no consistent effect was apparent" (p. 388).

Rater and rating scale.

With the shift from grammatical accuracy to communicative competence in Second Language Acquisition (SLA), rating has become much of a challenge. Raters in paired testing are generally confronted with three kinds of problems; the first of those is the inadequacy of rating scale used. Raters are often provided with rating scales that are mostly designed for individual performances, which they try to apply to paired performances. Unfortunately learner production in paired testing reveals many interaction patterns, few of which are described by existing rating scales. The second issue derives from the first one, and is related to the way raters try to cope with poor rating scales; studies show that for making their decisions about candidates' performances, rater often take into account take many aspects of candidates' performances which are not described by rating scales, such as non-verbal language used by candidates during their performance and supportive listening techniques and other communication devices used by candidates. The consequence is that since these aspects have not been formally described they are often perceived differently from one rater to the other (Ducasse & Brown, 2009). The third

issue is that even trained and experienced raters, as reported by May (2006), do not attend to the same features when rating paired performances even though they reach the same grades. Her work confirms Orr (2002), which shows that many raters have difficulty in adhering to the assessment criteria. According to Orr, "the verbal reports of many raters show difficulty in adhering to the assessment criteria. There is also evidence that raters do not understand the model of communicative language ability on which the rating scales are based" (p. 153). This finding is all the more disturbing that assessment criteria are key tools for reliability. These studies call for more rater training, and further examination into the rating scales, in order to ascertain whether adjustments need to be made. They also call for effort to find correct descriptors for assessing oral performances.

Learners' preferences.

The issue of learner preferences remains a controversial one according to the published literature (see Együd & Glover, 2001; Marochi, 2008; Taylor, 2001). Marochi (2008) concluded that the participants prefer the individual test. Indeed out of the 10 participants in her study eight (8) answered that they preferred the individual test. Her results contracted the results published by Együd and Glover (2001) and those of Baker (2000) reported by Taylor (2001). These two works concluded that learners generally prefer the paired format. Baker used a questionnaire with 130 candidates who had taken a variety of Cambridge EFL tests (PET, FCE, CAE, CPE, or CEIBT); all the respondents agreed or strongly agreed with the 12 statements in her questionnaire (Taylor, 2001, p.16). A similar result was published by Együd and Glover (2001); their data revealed that all the 14 participants liked the paired test.

Many reasons have been listed to support learners' preferences. The learners who preferred the paired format argued that it reduced their nervousness and that the paired test was

easier because they could receive some help from the partner (Együd & Glover, 2001). The participants who preferred the individual test, that is, the majority of students in the study of Marochi (2008), answered that they did not like the paired test format because they did not know their partners, so it was difficult to interact with them. Also, they declared that they were distracted by the partner because they did not know whether they should pay attention to their own production or what the partner says.

Research Questions

In nearly two decades of coexistence the two oral testing formats, the individual test and the paired test, have attracted many research studies. However there is still a shortage of published studies on candidates' preferences between paired testing and individual testing. The present study aims to contribute to fill that niche by studying the issue of preferences with ESL/EFL students. Three basic questions have guided the study:

1. Which of individual and paired testing do learners prefer?
2. Is there any correlation between learners' preferences and their performance on each test type?
3. What reasons underlie learners' choices?

Chapter 2 – Method

Qualitative Method

This research uses a qualitative approach; the choice of qualitative methods is dictated by the nature of the study and research objectives. Indeed, what learners feel or think when taking oral tests is a subjective experience; it varies from one person to the other. In order to access these feelings, to understand and describe them qualitative methods are considered as the most appropriate methods. According to Woods (2006) “The qualitative researcher seeks to discover the meanings that participants attach to their behavior, how they interpret situations, and what their perspectives are on particular issues” (p.4) This corresponds to the nature of the present study, which is concerned with choices participants make and their underlying reasons for these choices. Previous research has provided some compelling results about learners’ preferences and their reasons, yet the current study does not intend to test those arguments and categories. Rather it adopts an exploratory approach to understand and explore the issue in order to uncover potentially new arguments and reasons behind learners’ preferences. According to Mackey and Gass (2005), such research is better conducted through qualitative methods.

The main data collection technique is the interview, as opposed to the questionnaire, which is often used in similar research. The choice of the interview is justified by the fact that in contrast with the questionnaire, the interview does not propose predetermined categories and answers into which learners have to fit their views. Such an approach would ignore, underscore, or overestimate some of participants’ feelings and opinions. The interview is, comparatively, believed to offers more space to capture a wider range of reactions and feelings. That is why the interview is described as flexible data collection tool; allowing researchers to take into account the unique experience of most participants. (Mackey & Gass, 2005).

Participants

The 15 participants in the study were international students enrolled in various courses and programs at a large university in the midwestern United States. Their length of residence in the United States varied from six months to three years at the time of the study. They were all adults, males and females from various countries (see Table 1). The research population of the study was a sample of convenience; participants were volunteers recruited in their English classes, through direct contact with the researcher on campus or among students who contacted the researcher after reading recruitment flyers posted on campus.

The participants fell into two groups of proficiency levels; four (4) were advanced and 11 were intermediate level learners. Intermediate level participants were students currently enrolled as intermediate level students in the English Language Center (ELC) at the university or students who had just completed English language classes at intermediate level. Advanced level participants were students currently enrolled in graduate programs after completing pre-academic English language classes or students who had been admitted directly into graduate programs on the basis of their TOEFL scores. No further test was administered to estimate their levels since this variable was not controlled in the study. Beginner level students were excluded from the study because English was the medium for the tests and the interviews; therefore minimum language proficiency was necessary.

Table 1: Participants by gender and origin

Participant	Gender	Origin	Proficiency level	Length of residence in US
P1	Female	China	Advanced	6 months
P2	Female	China	Advanced	6 months
P3	Male	China	Intermediate	6 months
P4	Male	China	Intermediate	6 months
P5	Male	Saudi Arabia	Intermediate	6 months
P6	Male	Iraq	Intermediate	6 months
P7	Male	Saudi Arabia	Intermediate	6 months
P8	Male	Saudi Arabia	Intermediate	6 months
P9	Male	Saudi Arabia	Intermediate	6 months
P10	Female	Russia	Intermediate	1 year
P11	Female	Russia	Intermediate	1 year
P12	Female	Korea	Advanced	4 years
P13	Female	Ghana	Advanced	18 months
P14	Male	Saudi Arabia	Intermediate	6 months
P15	Male	Saudi Arabia	Intermediate	6 months

Instruments

Three types of data were collected from participants, which required two types of instruments. The first group of instruments consisted of oral testing tools; they were test prompts for individual and paired testing, and the analytical grading scales adapted for the tests. The second group of instruments consisted of the interview questions.

Oral testing tools.

Instruments for paired testing. The material used was a commercial copy for the preparation of the Cambridge ESOL FCE examination (University of Cambridge, 2008). Two pages of the book were copied from the book onto two A4 sheets and laminated to serve as speaking prompt in the paired testing. This material corresponds to that used in "part 3, two way

collaborative task" of the Cambridge FCE examination. This prompt showed seven independent pictures of (a) three happy girlfriends having a coffee, (b) a table covered with various dishes, (c) two lovers standing and watching the sunset, (d) the front side and the walkway of a little country house, (e) a grandfather and a grandmother reading to their two grandsons sitting on their laps, (f) a set of British coins, and (g) a beach resort with a couple of tourists.

Participants were asked to discuss the importance of these objects and their contribution to happiness and then choose the two most important for them. The pictures were accompanied by an examiner script to make sure the examiner treated all the participants equally. An example of an interlocutor frame and the accompanying visual are provided in Appendix A.

Instruments for individual testing. This part of the testing was a duplicate of Part 2 of Cambridge IELTS examination (Danesh, 2007). A topic card bearing a discussion topic was designed for candidates. The interviewer was provided with a script (appendix B). The content of the two tests were different to avoid a test retest effect that might otherwise make the second test easier, since both tests were taken consecutively.

Rating scales. The rating scale used in the study was adapted from the rating scale developed by Bonk and Ockey (2003). The purpose of the scale was to provide a clear idea of aspects of competence to focus on during participants' performance and ensure reliability (Douglass, 1994). Although the two oral exams had their own analytic rating scales I decided not to use them for the study because the two rating scales used different measures. IELTS uses a 10-band rating scale (IELTS, 2006) while the Cambridge FCE uses a 5-band scale (O'Sullivan, 2008). So the original rating scales of the two exams were discarded in favor of an alternative rating scale to ensure ease of comparison between grades obtained and the ease of use by

examiners. Instead, the rating scale developed by Bonk and Ockey was chosen because it proposed clear descriptors of speaking performances with the grades to attach to them, and also it is flexible enough to allow the rater some autonomy of judgment.

Two different rating scales were developed from the unique model of Bonk and Ockey (2003); one for the paired test format and the second one for the individual test format (see appendix D and E). The rating scale of Bonk and Ockey was developed to take into account interactive communication strategies, which are important features of the paired test format. To get the scale used for the paired test very few details were changed from the original scale. But to obtain the rating scale used for the individual test the entire fifth column, the c*ommunicative skills/strategies* column, was removed.

Interview guide.

A second set of data consists of interviews conducted with a semi-structured interview guide (see Appendix C). According to Merriam, (1998), “the rationale behind the use of interview as a data collection tool was that it can provide access to things that cannot be directly observed, such as feelings, thoughts, intentions, or beliefs” (cited in Ohata, 2005, p. 140). This view is consistent with the purpose of the study, which was to bring participants to voice out their feelings and thoughts about the two kinds of oral tests to which they had been exposed.

Procedure

Participant recruitment.

The data were collected from February 22 to March 18, 2011. Some participants in the study were first contacted directly in their English Language Center (ELC) classes after permission was obtained from the coordinator of ELC classes and the teachers. In three to five minutes students were informed about the goals of the research and participants' role, and then they were invited to enroll. They were also informed that all information would be kept confidential, and also that participation was voluntary. During the class visits learners were given the email address of the researcher so that those who were interested could write to enroll. This strategy was not successful; although many students showed enthusiasm for the research in the presence of the researcher and their class teacher no students wrote or visited the researcher to enroll. So the researcher had to change his approach when visiting subsequent new classes; during the visit the researcher asked students who were interested to give their email addresses. A blank sheet was sent round the class so that those who were interested could write down their email contacts. Next, the researcher sent invitation emails to those students. This last strategy proved more efficient, more than half of the participants were enrolled this way. Other participants answered after reading the posters that were posted on campus near classrooms. And finally some participants were recruited through direct contact in residence halls.

Data collection.

The data collection was conducted by the main researcher and a second researcher. Both researchers were adult males in their second year of Master of Arts in Teaching English to Speakers of Other Languages (MA TESOL). They were both international students with two

years of residence in the United States. Prior to enrolling on the MA TESOL program the main researcher and the second researcher were both teachers at the national teacher training college in their country of origin, Ivory Coast (West Africa), where they were involved in the training of future English teachers. They hold respectively a secondary school teacher's degree in English and a Doctorate obtained at the English Department. They total respectively 21 and 10 years of teaching and testing in EFL context. They had limited experience with teaching and testing in the USA; prior to the current research project they had taken part to the placement test for ESL student and taught those ESL students for one semester as part of their MA TESOL training program. They had also taken part to another placement test for the English Language Center (ELC) where both individual and paired oral test were used a part of the placement test battery. As training for the present research both researchers watched and rated a series of six videos of FCE and IELTs tests (three of each test) downloaded from the Internet. They both collaborated to the selection and adaptation of the rating scale used in the study

Before starting the tests participants were invited to read, discuss and sign consent forms. They were reminded again that they could suggest any modifications to the data collection or withdraw from the study at any time. No participant at this level decided to withdraw and they all gave their informed written consent for recording the tests and interviews.

The research had two main stages which had to be executed in strict order: (a) the testing stage, which consisted of two tests, and (b) the interview. The order of the administration of the two oral tests was not controlled, so participants took the paired test and the individual test according to the time they signed in for the study. When the two participants scheduled arrived together as many did, they started with the paired test. Otherwise the participants who arrived earlier took the individual test before the arrival of his or her partner.

Most participants knew each other before the test; they were classmates or friends. Seven participants were from the same class and they were paired at random according to their availability; six participants, that is, three pairs, were roommates or close friends; only two participants had never met before the study. Although it was not done intentionally, most pairs had almost the same proficiency level.

Individual and paired testing sessions. One or both researchers were involved either as assessors or interlocutors[1]. This arrangement was made necessary because of the paired testing component of the research; as prescribed by Cambridge ESOL the paired testing sessions required two examiners. The two researchers participated in the first stage of the data collection (individual and paired testing) either as assessor or interlocutor. assessor 1, the assistant researcher, was the interlocutor and he provided all the grades during individual tests. During paired testing sessions the second researcher was both the interlocutor and the first assessor. He provided holistic scores for participants' performances, while second assessor, the researcher, was the silent scorer using the analytic rating scale to score participants' performances.

Individual testing.

The individual testing component of the study was adapted from and conducted according to Part Two of the (IELTS) International English Language Testing System (IELTS, 2006). In this part participants had to produce a presentation of one to two minutes on a topic of general interest. Participants were provided with a topic card (a sheet of paper containing the topic that they had to discuss). The examiner read the topic with the participant to make sure the

[1] In the literature dealing with paired testing the two examiners are referred to as the *assessor* (the quiet examiner) and the *interlocutor*, who interacts with candidates (Foot, 1999; Saville & Hargreaves, 1999)

topic was well understood. Next the participant was given one minute to gather ideas before he or she started speaking. A blank sheet of paper was provided in case the participant might want to write down ideas. The participants' presentations lasted one to two minutes, after which the examiner sometimes asked a few questions to elicit more answers depending on the participant's presentation. After the presentation the examiner graded the performance using the rating scale provided.

Paired testing.

The paired testing was an example of the FCE test "Part 3, Two-way Collaborative Task," which involves the two test takers in a two-way discussion; the task consists entirely of peer–peer talk. The task was set up by the examiner according to the instructions specified in the *interlocutor frame*. The *interlocutor frame is* a script that specifies the language to be used by the examiner when introducing the tasks. An example of an interlocutor frame and the accompanying visual is provided in Appendix A. In execution of the interlocutor frame the interlocutor asked participants if they had understood the task. When they answered positively they were asked to begin the peer-to-peer discussion. Following their discussion the interlocutor (the second researcher) provided individual scores to each participant based on his global judgment of their performance. The second examiner (the main researcher) often sat at the far end of the table or of the room and used the analytic rater scale to assess participants' performances throughout their interaction. The score of each participant was the average of the two scores awarded by both raters.

Video recordings.

The second set of data collected from participants consisted of video recording of the test sessions. Each participant was recorded during the individual testing and the paired testing session. For both sessions the video data about each participant was about ten minutes long. The recordings were made with a digital Sony HD Handicam; DCR-SR68 mounted on a tripod. The video recording was planned to provide additional data and allow comparison with participants' interview data. First, the video images helped verify the reality of some information provided by participants during interviews. Secondly, the video recording provided additional data, as it helped capture some important attitudes about participants that the interview questions did not elicit.

Interviews.

All the interviews were recorded with a digital voice recorder; Sony, ICD-Px820. When transferred on the computer the result was audible enough to allow a comfortable listening and transcription. The interviews were all conducted in English by the main researcher. The average duration of interviews was about 15 minutes; the longest interview lasted 26 minutes and the shortest 12 minutes.

All the interviews took place immediately after the two tests, except for one participant whose interview was delayed till the next day to allow him to attend his next class. The interviews were conducted on a semi-structured basis. According to the description of Merriam (2004), "the semi structured interview contains a mix of more and less structured questions. […] The largest part of the interview is guided by a list of questions or issues to be explored, and neither the exact wording nor the order of the questions is determined ahead of time" (p.13).

Data analysis

Grades.

Each participant was awarded two grades; one grade on the individual test and the other for the paired test. The paired testing grades were obtained by running the average of the grades proposed by the two researchers for the performance of the participant (Saville & Hargreave 1999, p. 48). As prescribed by FCE the interviewer/interlocutor proposed a holistic grade while the scorer used the rating scale to propose a second grade. The inter-rater reliability between the two raters was verified by calculating the Pearson correlation ($r = .85$) for ($n=15$). The correlation was significant at the 0.01 level (2-tailed). Individual testing grades were all awarded by the second researcher to ensure consistency.

No research question was dependent on the grades alone. But they contributed together with interview data to answer research questions related to the correlation of participants' test preference and the grades they obtained during the tests. Therefore the grades were compared to interview answers related to participants' choices.

Videos.

Videos of the tests were not transcribed, but they were all reviewed to triangulate data provided by participants during interviews. The videos permitted confirmation of some of the themes mentioned by participants and allowed the researcher to make sense of some answers given by participants during interviews. The additional themes collected from video images were integrated to the themes recorded from the interviews and used as framework for analyzing interview data.

Interviews.

The interview data were analyzed through the inductive approach; in other words, "research findings emerged from the frequent, dominant, or significant themes inherent in raw data, without the restraints imposed by structured methodologies" (Thomas, 2006, p. 238). After data collection 11 out of the 15 interviews were transcribed completely. The rest of interview recordings were reviewed in their audio format and extracts of the significant parts were transcribed and used in the analysis. The transcribed interviews were revised and emerging themes highlighted. Similar themes were put together by "cut and paste" with a word processor. Next they were synthesized into categories according to research objectives. Finally, the categories emerged both from interview questions and from answers provided by respondent.

Chapter 3 – Findings

Test Scores

As indicated earlier in the method section, the highest score that a participant could obtain was 5, since performances were graded using a 5-point scale rating scale. Overall, there was no difference in the number of students who got better grades in individual test format and those who scored higher on the paired format as evidenced by their test scores. A tally of the participants' performances (see Table 4) revealed that an equal number of students (seven students) got higher scores on each test format over a total of 14 students (one student (P2) was not taken into account in the tally because that participant got the same score on both tests). The descriptive statistics of the average scores showed that the mean of the individual test scores (mean =3.55; SD .82628) was higher than the mean of the scores on the paired format (mean =3.4267; SD .72798). However there was no statistical significant difference in the score of individual tests and the score of the paired tests; $t(14) = -1.317$, $p > .005$.

Table 2: Descriptive statistics of participants' performance in the two test formats (N = 15)

Test format	N	Means	SD	Minimum	Maximum
Individual	15	3.5583	.82628	2	5
Paired	15	3.4267	.72798	2	4.90

A feature worth noting was that nearly all advanced participants (three out of four participants) scored higher on the individual test while most intermediate students scored higher in the paired test; that is 7 participants out of a total of 10 intermediate students (see table 4).

Table 3: Proficiency levels and scores

Participants	Proficiency Level	Paired test scores	Individual test scores
P1	**Advanced**	3.9	4.375
P2	**Advanced**	3.45	4.375
P3	Intermediate	3.55	4.25
P4	Intermediate	3.4	3.375
P5	Intermediate	2	2
P6	Intermediate	3.35	3.625
P7	Intermediate	3.35	3.25
P8	Intermediate	3.05	3.5
P9	Intermediate	3.25	3
P10	Intermediate	3.1	3
P11	Intermediate	3.2	3.5
P12	**Advanced**	4.85	4.625
P13	**Advanced**	4.9	5
P14	Intermediate	3.4	2.875
P15	Intermediate	2.65	2.625

Preferences and Correlation Between Preferences and Performances

This section addresses the first two research questions about the type of test format participants preferred and whether their preferences correlated with their test scores. Since the data were collected using a semi-structured interview scheme, participants' answers about the test format they preferred did not come from one single question. Often, participants provided elements while answering other questions; as a result some respondents proposed their opinion more than once. But for confirmation, a specific question was set toward the end of the interview to allow participants to clearly express their preferences. The answer to the first research question was therefore induced from several comments and the question which read: *For a future oral test which of these two test types would you rather have?*

The data show that 11 participants declared that they liked the individual test, so if they had the choice for a future oral proficiency test they would rather have it. Four participants

expressed their preferences for the paired test format. So in total eleven out of fifteen (11/15) participants in the study expressed preference for the paired test. In other words, the majority of the participants in the study preferred the individual testing.

Participants' preferences were collected while they had no knowledge of their grades. Results show that the preferences expressed by participants did not always correspond to the test format type where they got higher grades as shown in *Table 4* .

Table 4: Participants' preferences and their scores

Participant code	Paired testing scores	Individual testing scores	Participants' preferences
P1	3.9	**4.375**	Paired testing
P2	3.45	**4.375**	Paired testing
P3	3.55	**4.25**	Individual testing
P4	**3.4**	3.375	Individual testing
P5	2	2	Individual testing
P6	3.35	**3.625**	Paired testing
P7	**3.35**	3.25	Individual testing
P8	3.05	**3.5**	Individual testing
P9	**3.25**	3	Individual testing
P10	**3.1**	3	Paired testing
P11	3.2	**3.5**	Individual testing
P12	**4.85**	4.625	Individual testing
P13	4.9	**5**	Individual testing
P14	**3.4**	2.875	Individual testing
P15	**2.65**	2.625	Individual testing

* The highest score of each participant is in bold type.

The data related to the correlation between participants' preferences and their grades (Table 4) show that four (4) participants in the study expressed preferences for the paired test format and half of those participants (P2 and P10) scored higher in paired testing. The other two participants (P1 and P11) received higher scores on the individual test. Among the 11 participants who preferred the individual test, five (5) scored higher in the individual test. In total there are only six cases in which participants' choices corresponded to the test where examiners awarded better grades. In sum, the analysis showed that in most cases there was no connection between a participant's preference and his or her actual scores on the tests. This result suggests that the actual grades do not play a major role in the choices made by participants. Not only participants' scores did not seem to guide the choices they made, but there is also no evidence that participants' preferences affected their grades either. In other words learners' decision to choose the paired test or the individual test format seemed entirely dependent on factors others than the grades.

For further investigation, participant's preferences were compared to their opinions about their performances. During the interview participants were asked to tell the test where they believed they had performed better and therefore where they expected better grades. This comparison gave a completely different result from the first one; in the majority of cases there was a clear connection between the test a participant preferred and the test where that participant expected a better grade. The choices of 12 participants out of 15 matched perfectly with the test format in which they predicted better grades (see Table 5).

Table 5: Participants test preferences and their highest score expectation

Participant codes	Expectations/predictions	Participants' preferences
P1	**Individual testing**	**Paired testing**
P2	Paired testing	Paired testing
P3	Individual testing	Individual testing
P4	**Paired testing**	**Individual testing**
P5	Individual testing	Individual testing
P6	Paired testing	Paired testing
P7	Individual testing	Individual testing
P8	Individual testing	Individual testing
P9	Individual testing	Individual testing
P10	**individual testing**	**Paired testing**
P11	Individual testing	Individual testing
P12	Individual testing	Individual testing
P13	Individual testing	Individual testing
P14	Individual testing	Individual testing
P15	Individual testing	Individual testing

* Diverging pairs are in bold type

The Reasons Behind Participants' Choices

The investigation of participants' choices and their grade expectations indicated that the choices made by the majority of participants were primarily guided by the fact that they expected better grades in the tests they chose. In other words they chose the test which they believed allowed them to get at least a pass score. However the question remained and could be rephrased as follows: What in the tests that participants chose made them believe they got or they could get a better grade? The interview data suggest that two main reasons that guided participants'

choices regarding the type of test format they preferred were the ease of the test and participants' level of anxiety during the tests. Participants' reasons and their sub-categories are summarized hereafter in *table 6*

Table 6: Summary of participants' arguments

Ease of test

1. Pictures
    - Pictures provide the content
    - Pictures are attractive
    - Pictures promote interaction
2. Collaborative and interactive task
    - The partner helps with ideas and language
    - Partners allow spontaneous communication
3. Test items and test specifications
    - The topic is easy
    - There is no distraction from a partner
    - There is plenty of time for speaking
    - The preparation period is helpful

Test anxiety

1. A non-threatening environment
    - The presence of a partner is comforting
    - The partner has a comparable language level
2. The examiner as a facilitator
    - Participants speak to a good listener
    - The examiner asks precise questions
    - The examiner does not judge candidates
    - Participants get feedback from the examiner

Ease of test.

Pictures.

Irrespective of the side taken many participants argued that the test they chose was the easiest. Many participants who chose the paired format as their preferred test, commented that the test was made easy by the pictures. As shown in *Table 7* this argument represents 12.5% of the total arguments presented by participants to justify their choices. As described earlier in the Method chapter, the speaking prompts in the paired test consisted of a series of seven pictures. Some of the participants declared that the pictures made the test easy because participants were free of the burden of creating the content of their discussion. On this issue some participants offered the following comments:

P6: The first one you have a picture, the picture describes what you wanted to say.

Other participants added that the pictures were appropriate for creating the required interaction between the partners.

P2: You give us lots of pictures and the second is I can have lots of interactions with my partners.

P1: (…) I like it because we have some pictures we can share. Sharing something with a friend it's not always that you have the same opinions. But it's something like information exchange.

With the pictures as input, in the paired test, the test looked simple and friendly to participants; they commented that they did not have to think much before finding ideas to express themselves, so they could concentrate on the interactions with their partners. The impact

of the pictures combined with other elements of the paired testing, such as the co-construction of the speech contributed to give the feeling that the paired test was an easier test.

Collaborative and interactive work.

Participants also said that the paired test was made easier by the collaboration with the partner. The participants who chose paired testing were happy because they did not have to do everything alone. In contrast with individual testing, participants could support each other with ideas and language. Many participants saw the paired test as an opportunity for collaborative work. Defining collaborative work Brown (2007) notes that "students work together in pairs and groups, they share information and come to each other's aid. They are a "team" whose players must work together in order to achieve goals successfully" (p. 53). Thus many participants did not see the test as a competition; they rather took advantage of their partner's production to improve their own production. For example, one participant (P6) was convinced that he could not have spoken on the individual test as much as he did on the paired test, because he would have fallen short of ideas and motivation. He declared:

> P6: The second [individual test] I have no energy, there is nothing that gives you energy to talking. In the first one there is many thing to give you energy. Like the discuss, like the image. But the second one just you talk alone maybe you're tired. And when I talk and I'm tired I have finished. But when, in the first one when I talk with my friend maybe when I'm tired he gives me advice or idea and I go with another idea. And there's a discuss and the discuss give more energy than the second one.

During the paired test some participants did more than just copying ideas and responding to cues from their partners. In their discussion participants P10 and P11, who shared the same first language often asked one another for translation of words; one of the participants sought help from the partner with the words *entertainment* and *driveway*. Other participants sought confirmation of the content of their ideas before speaking out.

P13: Do you think these are their children or grandchildren?

P12: Grandchildren

P13: Grandchildren, yea, yea. And sometimes grandchildren too can bring a lot of happiness…

Some participants believed that during the paired testing they had a real purpose for talking. They considered that they had an audience which they wanted to convince, and this motivated them to find ideas and speak as much as necessary to make their points. On the individual test these participants felt things differently; they felt they were talking for an examiner who was interested not in their ideas but in only the quality of their language.

P1: For me I think (…) the examiner don't want to know what I have to say. He wants to know my language ability not my opinion.

Interviewer: So you think that the individual test is about your language ability. And what about the pair test?

P1: it's about interaction and your behavior.

For this reason the participants enjoyed talking to the partner who appeared as a real conversation partner; sometimes helping them through his or her questions and opinions. The

participants argued that the presence of the partner was even more helpful when the test was difficult. This is what this participant declared:

P2: If it is a little harder, I think we can perform better with our partners.

Participant P2 believed that the partner could be helpful no matter his or her level of competence. If the partner is more proficient he or she could be used as a resource person. If such a situation happened the participant suggested:

P2: I can learn from her, I can imitate it. Why not?

According to participants the partner played an important role. Thanks to the partner the task became easier because they were not alone to produce the ideas and the language; the final product was produced collaboratively. They could take advantage of the ideas and language produced by the other partner to improve their own performance. As a result participants could perform above their actual level by imitating, copying, or taking cues from what the other partner says.

Test items and test specifications.

Like the participants who expressed preference for the paired test, some participants who chose the individual test also based their choices on the ease of the test. On the second exam participants were asked to prepare and present a short talk whose prompt was: *"I would like you to tell me about the job you wanted to do when you were a child"* (see appendix C). They believed that the test was easier than the paired test thanks to the exam question and time specifications. For them, the first factor making the individual test easier was the topic.

Although many things could have contributed to the comprehension of the topic, the participants answered that they found the exam prompt for the individual test easy, that they understood exactly the exam prompt so they had a clear idea of what was required from them during this exam. They could therefore work serenely to achieve those objectives. This conviction was perceptible through their attitudes; video images of the test sessions show that when participants were invited by the examiner to start speaking they did so without hesitation. On the other hand one could notice lots of false starts, and whispering at the beginning of paired testing sessions.

In the understanding of the participants the individual test was an opportunity for expressing themselves on a topic, to provide enough evidence for the examiner to assess the quality of their language. Therefore understanding the topic was very important, because it conditioned the amount of language they could produce in the time that was allotted. Here are some comments that the participants made on the topic:

P13: I'm happy about the topic itself, and the way the examiner presented it to me the clues that he gave me helped me have a good thinking about the question, and the question was pretty easy. The words were not difficult and the sentences were not too long that might confuse me or something. So because of these elements I was able to (.)

P13: I think the first one I may get a good grade. Because it was not all that difficult, and it was just a way of life. It is about life and it is about something

that I've gone through before. So it is not like something that I have to go through deep thinking so I think it's okay.

P12: He didn't ask me anything to analyze, so like that just express my own impressions. For example in my class when I have to offer my own analysis of some matter, I find it much harder. To talk about than when I'm just speaking with friend about what I think…

For many participants the topic did not pose much intellectual challenge. Under those circumstances most of participants who chose individual testing believed that they had performed well and would get better grades on the individual test than they would on the paired test. Through the words of participants the topic was easy because it met some conditions: it dealt with a life issue; it was phrased with easy words, it was short, and it did not ask to analyze a situation but simply to narrate an event that they had lived personally.

The participants also found the topic easy because it allowed the use of some exam strategies for getting a good grade. Some participants commented that on individual testing it was possible to perform well and possibly get good grades by using different exam strategies like avoidance and memorization. They revealed that the topics of the individual testing made it possible for them to avoid some the difficulties, like the shortage of ideas and complex language forms. In the following extract one participant explained his strategy for giving a good performance by avoiding the troublesome aspects of the topic.

P4: You can think about it and choose a better one for you to say. Like I just said I wanted I to be a tour guide. Maybe in my childhood I didn't think about tour guide, but I think tour guide is easier to discuss

In the study participants were asked to discuss the jobs that they wanted to do in their childhood, participant P3 answered that he wanted to be a tour guide and he built his presentation on that job. But during the interview he revealed that in his childhood he had never thought of becoming a tour guide. He discussed this job during the test because he had found it was easy and convenient; he had travelled to different countries so he had the content and the necessary language to discuss the profession of tour guide. He avoided discussing a career in the army that he was really dreaming of in his childhood because during the test he was no feeling confident enough to tackle this topic. Such an exam strategy was rather difficult to apply with paired testing where the content was imposed through the pictures.

The other strategy used was memorization; some Chinese participants in the study claimed that Chinese students have become so familiar with the topics and the individual test format that they believed this exam was no longer appropriate for assessing the real level of Chinese students. Here is an extract where a Chinese participant revealed that candidates could get better grades by memorizing some types of topics.

P1: We have a list of the hot topics for individual [test]. So many students study and memorize them, and they speak out, they recite. So the individual test is better for Chinese people. You can't test their real speaking level. They can recite. It's a big problem. But in the pair one you must say something. It depends on the condition of your partner.

This participant's comments suggest that many Chinese might prefer the individual test because it would help them pass their test easily. Yet the participant (P1) preferred the paired test because it looked like a better test, a more valid test of candidates' oral proficiency. For her this test created the conditions for candidate to produce spontaneous discourse.

Many participants said that they preferred the individual test because the topic allowed them to use several techniques to have good grades without being really competent in the language. Candidates could have good grades by drawing on their experience with exams and a good preparation. By contrast, the topic of the paired test did not allow much freedom and it was difficult to prepare everything beforehand. Also, even the best preparation could be upset because the partner may bring in new issues.

In addition to the topic, the time specification of the individual test was believed also to be helpful. In their comments some participants revealed that the individual test was also made easy by the preparation time allowed by the test. The preparation period which was awarded at the beginning of the test was seen as essential for a good performance and a good grade. As many participants remarked, they found the paired testing difficult because they did not have such period for planning their interaction. To illustrate this opinion one participant proposed the following idea for improving the paired testing:

> P1**:** I think a more clear instruction or a communication with your partner before we take the video.

For this participant the paired testing was difficult because there was no preparation time when partners could discuss the ideas that they would present in front of the examiner. Participants believed that the preparation time contributed largely to their good performance

during the individual test. When questioned about the test where they believed they had done better some of the participants who chose the individual test proposed the following answers were proposed:

P4: Because for the second [the individual][2] one you have some time to prepare it, like one minute to prepare it. You can write down and read it.

P11: In second [individual test] test because I had the time to prepare, I wrote some notes and the examiner listened for me without stopping me during my speech, without showing my mistakes.

P11: I prefer the second test because, as I said, I had had a time to prepare and it's very important for me. It's hard for me to speak without prepare

The participants agreed that the preparation time was very beneficial. However their comments showed that they used the preparation differently.

P7: Because you can prepare your ideas in your mind and make steps how you can achieve the conclusions for each idea. For example you want to talk about, information about the first idea and make connection with the other idea. So it is easier to talk about and prepare some information and talk about them, and you know I'll talk about his idea, the second idea.

P4 Yes of course, but for the second test you can write down some things so you don't need to worry about grammar and vocabulary. What you need to worry about is clear, make your speak clear.

[2] Participants took the tests in different orders. Some took the individual test as a first test, when others took it as a second test.

Other participants did not share this positive opinion about the preparation time. They some participants felt that the preparation time made the individual test less realistic. This feeling was shared by many participants who chose paired testing. Even though some participants, like P1, were convinced that they performed better on the individual test, they expressed preference for the paired test because it looked more appropriate for assessing learners' real communicative skills. She made the following statements:

P1: The individual test is better for Chinese people. But it can't test their real speaking level. They can recite. It's a big problem. But in the pair one you must say something. It depends on the condition of your partner.

One problem is that in the individual one, we have one minute to prepare. So my partner went like; the first reason, the second reason, the third reason. So I don't think it is a real way of communicating, a way of talking in daily life. It's not very natural.

For this participant, the language produced on the individual test was not much different from the written language. She blamed this situation on the preparation time which was awarded to participants before they start talking. The same point was also made by participant P4:

Interviewer: When you compare both tests, in which test do you think you produced better English?

P4: I think the first one [paired test]. Because for the second one you have some time to prepare it, like one minute to prepare it. You can write down and read

it. You don't need to speak it. For the second one, when the other guy says something I have to response. So I think this will prove my English better.

Despite these negative reactions about the drawbacks of the preparation time, mainly recorded among those who preferred the paired format, there was a general agreement on the fact that the preparation time was a very helpful element for those participants who based their choices on the grade prospects and the ease of the test. Not only may it have contributed to the understanding of the topic, but some participants believed that the preparation time also helped with their performances, although this belief is not borne out by test scores data.

Participants also made remarks on the issue of the partner in relation with the individual test. Since they were alone in the individual test participants commented that they were not disturbed during their performance by another candidate, so they were coherent and spoke more than in the paired test. They did not suffer any distraction, or interruption from a partner that might make them loose track of their ideas. So they performed in a comfortable atmosphere, they could develop their ideas and make the points they wanted. They declared that the discussion with the partner and the questions of the partner in the paired test did not help them speak as they wanted to. In other words they felt that the individual test was free from all the sources of distraction present in the paired testing. This point is illustrated by the following extract:

P4: In the second test I can say whatever I want, I don't need to think about his opinion. (..) I express my own opinion.

Indeed some participants felt sorry that they had to sacrifice coherence in their ideas and even sometimes they had to give up their ideas entirely, in order to accommodate the partner who did not seem to understand the topic correctly or to follow the partner who was developing a

different line of thought. In contrast, participants believed that they had a double advantage with the individual test; first they had enough time to brainstorm and plan the content of their performances, and secondly when performing on the individual test they could say entirely what they had planned. Therefore, participants ended the paired test with the feeling that they had not been able to convince the examiner because they could not develop their arguments completely and coherently.

P1: (…) I am a person who will change according to another person's behavior. So in the first one I was ready to discuss all the pictures, but she [partner] described only two. So I picked one of her pictures, as my favorite and described it. And ignore the rest.

P7: Because you can prepare your ideas in your mind and make steps how you can achieve the conclusions for each idea. For example you want to talk about, information about the first idea and make connection with the other idea. So it is easier to talk about and prepare some information and talk about them, and you know I'll talk about his idea, the second idea. But with partner sometimes, he mixes your ideas, the second idea suddenly will become the fifth, so you can't very relax with this.

P7: because, before I can imagine, I will talk about this, and this, I make a schedule in my mind, I will talk about his, this idea, and this idea and so on. Not like with my partner. I talk about this idea, and want to enter into this idea, and suddenly he changes all the composition. The continuous speaking, something drops with the partners.

As a result of being the sole speakers in the individual test, participants believed they had enough time to express themselves; they made good use of the 2 to 3 minutes they had for their talk and they were happy that they were not interrupted during their performances. Participants enjoyed their experience with the individual test and they wished the paired test could be improved to provide them with similar comfort. Therefore when they were asked to suggest ways for improving the paired test, participants proposed that the examiner should monitor speaking time so that each candidate gets enough time to show his or her abilities; or to increase the time of the paired testing, to make it at least twice as long as the time of the individual test.

P3: maybe, I think to make some requirements; some people have to say for 5 minutes or 6 minutes and everybody has to say something.

P13: I don't know, I didn't time the testing times, but I think the two tests had about the same time, that's what I think, so if …. We are two then I think the time should be doubled.

P7: There is a difference between partners in language and some partners are not very helpful. They are selfish; they try to dominate the conversation, and send a message to the examiner 'I'm the best, I' the hero, and don't care about the other.

Test anxiety.

A non-threatening environment.

The second category of arguments behind the preferences expressed by participants was related to how comfortable they felt during the test. Three out of four (3/4) participants who chose the paired format declared that they felt very little or no stress during the paired test. They justified this state of comfort, first of all, by the fact that they were speaking directly to a partner, a friend for the most part, rather than addressing the examiner who was an adult with whom they were not very familiar. For those participants, in the paired testing situation they did not have the impression of taking a test, but they had the impression of holding a simple conversation with a friend. So they described the testing session as “fun” and “happiness” (P3).

Interviewer: How were you feeling in the first and the second test?

P4: More fun in the first test [paired test].

Interviewer: Excuse me

P4: More fun, more happiness in the test.

Many participants also believed that the paired test decreased nervousness; the presence of the partner helped them defuse exam stress.

P10: I don’t know it is just psychology, I said the person is not a problem for me, but when it is exam [the examiner] is not just a person but the person who judges me. Maybe, it is like a policeman so I prefer my friend to be here.

The paired testing was also said to reduce nervousness because it involved two partners of the same language level.

P2: I think we can perform better with our partners. Because we will have the same level of English study and we will feel less nervous.

This participant also declared that she would feel more confident if the partner had a language level lower than hers. For these participants there was definitely an advantage in having an interlocutor other than the examiner during the test. It must be the case that the participants who were not always confident with their speaking skills preferred not to interact with a person, the examiner, whose language level was higher and intimidating; Under such circumstances the partner appeared as the ideal interlocutor. By his or her presence the partner also reduced the exam stress and helped achieve a good performance.

The examiner as facilitator.

The participants who chose the individual test also believed that it was a better test because they felt at ease during this test. They believed that the individual test caused less anxiety and nervousness. Many declared that they did not feel much stress or tension during the exam. Their relative ease, compared to the way they felt during the paired testing stemmed first, from the differential effect of the partner and the examiner on them, secondly from their own personalities and attitudes in the presence of one or more people, and finally a feeling of insecurity regarding the confidentiality of their performance. In sum, they saw the examiner as a sort of catalyst, a facilitator, or a supportive listener that they could trust.

Firstly, many of the participants in the group answered that they felt less nervous while performing in front of the examiner, because the examiner was relaxed, smiling and sometimes supportive. They felt that he created the right environment and atmosphere for a good performance.

P1: A very relax atmosphere. You know many of my friends; they can't speak to the examiner. They look at the examiner in the house and they say "oh I'm very nervous". So a very kind and smiling examiner will make you relax.

P12: I felt calm, if I had to decide I would say I enjoyed talking to the examiner a bit more, because he was outgoing. This lady [her partner] was also pleasant to talk to she was just a bit calmer. But it is a matter of personal preference I guess, I prefer to talk to people who will respond enthusiastically That's about 50% of the reason why I enjoyed talking to the examiner a bit more.

The absence of such a person with good listener's attitudes on the paired test was felt as a big handicap by participants. Video images of test sessions showed that even on paired testing sessions some participants were talking to the examiner, seeking his approval on what they were saying instead of facing and talking to their partner as would normally do with a conversation partner. The listener issue appeared in many comments and was, for some participants, a deciding argument in favor of the individual test.

Ignorance of the way the paired test was scored often made participants express preference for the individual test. Participants wrongly believed that while working in pairs their individual grades depended on the performance of their partner or on the quality of their joint production. As a consequence of this feeling participants appeared tense during the paired test,

they spent the time worrying about their partner rather than focusing on their own performances or the task. Videos of the paired testing sessions show that after he realized that he has almost been doing all the speaking, the participant P9 started interviewing his partner to give him the opportunity to talk. At the opposite end, during the individual testing participants felt free and they displayed the best of their speaking skills. Video images (of their individual tests) often showed them more focused, and busy trying to communicate their ideas.

The state of nervousness became even worse when, in the course of the test, participants felt that their partners' level of language was not sufficient to help accomplish the common task successfully. During the interview some participants mentioned their partner's lack of experience in question asking, language errors, difficulties in answering questions, and finally the lack of fluency for a successful conversation (P9; P7; P8). Such partners were a source of concern and anxiety during paired testing. Therefore participants preferred the relative comfort they felt speaking with the examiner who spoke clearly and correctly. In the absence of an examiner they would rather talk to a partner with higher proficiency level. Here are some extracts of related comments:

> P9: That usually make some students nervous, when you talk to a partner who can't speak English very well, he can't speak English very well, when his questions are not very clear, and they can't respond my questions, it makes me nervous and. There are so many in this part. But when I take it with a speaker who is very well in English language that maybe I think it will help me. (…) because he can ask me the right questions and he doesn't make me nervous me.

P7: sometimes there is difference in level, language level, sometimes he takes long time to respond to your questions or your conversation and sometimes he is not clear as sound or ideas. And also some of them are selfish they try to dominate the conversation and send a message to the examiner, “I’m the best, I’m the hero, and don’t care about the other.

P7: I prefer higher than me because higher it is still, I can understand him and maybe I can imagine what he will do his next sentences… But lower than me I can’t imagine what they want to say. His voice not clear, his sentences are sometimes not said in correct way verb subject switch to hard to understand.

P9: I know when I took the test [individual test] … I can understand what he [the examiner] talked about. And he asked me clear questions and he knows what he asks me about. Sometimes the partner ask me question with incorrect grammar and I can’t understand. Because the partner doesn’t have experience for ask questions

P1: For the individual one I was free to say everything.
In the first one you must... It is like, sometimes in the daily conversation you worry about if I can understand my partner and if my partner can understand me.

A participant who described herself as “shy” or “super cautious” (P12) also opted for the individual test because she felt at ease in that test. She justified her attitude by the fact that prefers to perform in front of a reduced number of people.

P12: If there are more people around me when I speak, so find myself getting very nervous, and I stumble more and make more mistakes. So when I was with just one person I was less nervous and talked better (…). When I'm speaking with less people, when I am calm, I find myself speaking with ease. I have a big ego so when I make mistakes in front of some people I feel really bad. So I'm afraid of making mistakes when I'm speaking in front of several people. But when I'm with I so I tend to speak less. But When I'm with only one person, even if (.) even when I make mistakes I'm more relax because there is only one person who knows about my mistake. So I think it helped.

For this participant her preference for the individual test came from the feeling that she might be judged and sometimes "mocked" (P 12) by their partners. She was afraid of the partner's reactions in case she made mistakes or if her performance was not up to the standard. For these reasons she preferred the test with the examiner. Here is an extract of her comment:

P12: It's much better for students, because students will feel less pressure of making mistakes like, they will be less afraid of making mistakes, because like there are fewer people who will know about their defects. And I know it because, like I'm taking two languages now, French and Japanese, and we have oral test in each course. And when I have to speak with a partner in front of the teacher, I'm like oh if I make mistakes she will know, like, how foolish I would sound, and also the teacher. But when I'm taking an individual test it would be just the teacher and me and she understands. Because, like, she understands that we're just learning the language and we

can make mistakes. So I feel less pressure than like when I'm speaking with a lot of people.

Finally some participants liked the individual test because they could tell during the test how well they were performing. They interpreted the examiner's reactions as positive or negative feedback which they used to shape their oral production. They knew that they could continue on the same path if the examiner looked pleased, and that they should refine their language or ideas if the examiner looked less pleased. One of the participants answered that she knew that she was doing well by reading on the face of the examiner.

Interviewer: How about your performance, how do you know that you performed well?

P13: because I saw it on the face of the examiner

When at the end of the interview this participants was asked to say what test she preferred, she opted for individual testing because she believed she performed better on that test based on the reactions of the examiner. Like her, many participants felt encouraged by the examiners' reactions (nods, back channels, laughs, and facial expressions), they tried to decode those attitudes from their examiner to know how well they were performing, and possibly improve on their ideas and language when they believed the examiner was not happy with their performances.

These participants complained that in paired testing they did not have any sign from the partner telling them whether what they were saying was correct or not. And in most cases they confessed that they did not trust their partner who had the same or lower level. Therefore they preferred the individual test where they were facing an examiner they could trust. During their

performances they tried to get feedback from him or they interpreted his attitudes, reactions, and all nonverbal clues. One partner argued that it is motivating for her to have a speech partner who responds "enthusiastically" to her conversation. Here is what she declared:

> P12: I felt calm, if I had to decide I would say I enjoyed talking to the examiner a bit more, because he was outgoing. This lady [her partner] was also pleasant to talk to she was just a bit calmer. But it is a matter of personal preference I guess, I prefer to talk to people who will respond enthusiastically That's about 50% of the reason why I enjoyed talking to the examiner a bit more.

Distribution of participants' arguments

Table 7 represents the distribution of the arguments that emerged from the analysis of participants' answers. Since not all the interviews were transcribed, this analysis concerns the subset of 11 interviews that were completely transcribed. Overall, roughly 2/3 of the arguments provided by participants to justify their test preferences were related to the *ease of test*.

As shown in Table 7 the reasons behind participants' preferences were discussed about 112 times in the subset analyzed. To tally occurrences, arguments were counted when they were mentioned for the first time in a response and any time they were repeated or rephrased for answering a new question; that is, they were not counted again if they were repeated in the same response.

According to the data the greatest number of arguments appeared under the category *ease of test*. A total of 69 reasons out of the 112 arguments appeared under this category; that is, 61.60% of reasons. Within this category, *ease of test*, the reasons that were more frequently

discussed by participants were first, the partner's assistance with language and ideas during paired tests; this argument was mentioned 10 times, which represents 8.92%. It was often discussed, as mentioned earlier, by the participants who preferred the paired test. On the other hand the participants who preferred the individual test often evoked arguments about the absence of distraction from the partner during individual test. This argument had the highest rate of appearance within this category; it was mentioned 18 times, that is 16.07 %.

The arguments related to *test anxiety* appeared 43 times, that is 38.40%. Under this category participants mostly mentioned the comforting role of the partner. It was argued that the presence of a familiar person or a partner with comparable language level reduced their nervousness. This argument appeared 14 times, that is, 12% in the subset analyzed.

Table 7: Distribution of participants' arguments

Participants' arguments	Frequency	%
Ease of test	**69**	61.61
1. Pictures	14	12.50
• *Pictures provide the content*	*4*	*3.57*
• *Pictures are attractive*	*5*	*4.46*
• *Pictures promote interaction*	*5*	*4.46*
2. Collaborative and interactive task	**17**	**15.18**
• *The partner helps with ideas and language*	*10*	*8.93*
• *Partners allow spontaneous communication*	*7*	*6.25*
3. Test items and test specifications	**38**	**33.93**
• *The topic is easy*	*7*	*6.25*
• *There is no distraction from a partner*	*18*	*16.07*
• *There is plenty of time for speaking*	*6*	*5.36*
• *The preparation period is helpful*	*7*	*6.25*
Test anxiety	**43**	**38.39**
1. A non threatening environment	22	19.64
• *The presence of a partner is comforting*	*14*	*12.50*
• *The partner has a comparable language level*	*8*	*7.14*
2. The examiner as a facilitator	**21**	**18.75**
• *Participants speak to a good listener*	*6*	*5.36*
• *The examiner asks precise questions*	*7*	*6.25*
• *The examiner does not judge candidates*	*5*	*4.46*
• *Participants get feedback from the examiner*	*3*	*2.68*
Total	112	100

Summary

Chapter three has discussed the findings of the study; the data show that in general participants were not unanimous about the type of testing they prefer; both types of testing have their supporters. Still the study reports a larger number of participants in favor of individual testing. A total of 11 participants liked the individual test out of the 15 participants of the study. The analysis of the data, in view of revealing a possible correlation between the testing format elected by participants and their grades, showed that in many cases there was no connection between a participant's preference and his or her actual scores on the tests. Yet, it was noticed that the choices expressed by participants corresponded almost perfectly to the test where they believed they had performed better. Finally the investigation about the reasons behind learners' preferences showed that many participants who favored the paired test provided arguments which valued the support they received from their partner, and the way the paired test promoted communication between the candidates. On the other hand, the other participants felt that presence and performance of their partner hurt their own performance; so they would rather talk to an examiner on the individual test. Also those participants preferred the individual test because they believed that they had more chances of winning better grades with the type of topic used in the individual test.

Chapter 4 – Discussion and Conclusion

Discussion

The study of the scores awarded calls attention to two issues. The first of those was that nearly all advanced level participants got higher grades in the individual test while most intermediate level participants (the less proficient learners) got their highest grades in the paired test format. This first observation raises a question regarding whether there is a relationship between the test format and the level of the learner. In other words do participants perform differently on the test (individual or paired test) according to their proficiency level? To the best of my knowledge such this issue has not yet been discussed by literature. However the data from this study seems to confirm the findings of previous studies dealing with the performances of participants working in pairs. Those studies (Brooks, 2009; Davis, 2009; Galaczi, 2008; Marochi, 2008), analyzed the interactions between speakers with different levels and concluded that there was a significant improvement of the performance of lower level speakers when working in pairs.

The second issue with the scores comes from the comparison between the scores from the individual test and those from the paired test. Contrary to Brooks (2009) who noted that "overall, the students' performance was better in the paired format" (p.350), the current study did not find evidence supporting this claim. The data of the study did not reveal any significant difference between the scores of the paired test and the grades of the individual test. It might be the case that the 5-band scale used in this study, instead of the 11-band scale used by Brooks, was too tight to reveal a real difference between performances.

On learners' preferences the results of the study indicate that the vast majority of EFL/ESL learners who took part in this study (11/15 or 73.33%) preferred the individual test to the paired test. The results are similar to Marochi's (2008) findings where the majority of candidates (8/10 or 80%) expressed preference for the individual test after taking both tests. But the results contradict those of Együd and Glover (2001), who reported that all of their fourteen participants liked the paired test format (14/14 or 100%). The difference between these two groups of studies may come from the objectives and methods of these studies. The present study and the relevant sections of the study of Marochi (2008) asked participants to elect the type of testing they preferred after taking both exam formats; this may explain why both studies have comparable results. On the other hand the method and purpose were not so clear in the study of Együd and Glover (2001). Apparently in their study participants were only asked to give their opinion about their experience with paired testing. The study did not include any comparison of participants' experience on the individual test and the paired test. The candidates were interviewed after taking only the paired test format.

The present study provides evidence supporting most reasons proposed by participants in previous research studies to justify their preferences. But new reasons were also mentioned by the participants in the current study. All participants' reasons were classified under two categories, ease of test and test anxiety (*table 6*). Depending on the side taken participants noticed different elements from the tests or they interpreted the same elements differently. Thus, items such as the pictures used with the paired test or the exam question used for the individual test were seen differently; they were seen as positive and capable of triggering language production by some participants, while they were seen as handicaps by the other side. Because of the conviction that their performance was judged on the amount of language produced some

participants felt unsecure with the individual test because they felt that they may not have enough to say. Similar argument has been suggested in (Marochi (2008). On the other hand other participants believed that the pictures used with the paired test imposed a strict language framework in which they are not sure to survive. A similar contrast was noted about the preparation time, which was a key argument in the choice made by many participants. Many participants who favored the individual test argued that it allowed them to prepare a coherent oral production. Yet some participants who had a preference for the paired test argued that it made the discourse less spontaneous and less natural.

The arguments about test anxiety represented another example of the differences in the interpretation of the same elements. Many participants answered that when they performed with a partner they felt less stressed because they were not alone in front of the examiner. But other participants did not like performing with a partner because he or she made them nervous.

Some of the reasons proposed by participants raise issues about partner characteristics. Partner characteristics include factors such as age, gender, familiarity, language ability, culture, etc., but the participants in this study referred mainly to the partner's language ability, an argument also mentioned by Foot (1999) against paired testing. Foot was concerned with the potential problem of pairing candidates with differing spoken proficiency. He noted that "unless the candidates are well-matched, their attempts to sustain a discussion are likely to be, and often are, faltering and desultory, and the outcome, for them a sense of frustration rather than of achievement" (p. 40). Many participants in the current study expressed their frustration in front of the difficulty they met with their partners; they estimated that their partners did not have the appropriate language ability for the successful accomplishment of the task. Although no

significant impact was found on grades, the data show that for many participants the frustrations they felt during the interactions are strong enough to make them opt for individual testing.

Some arguments provided by participants to justify their choices touched on the face validity of the tests. One participant expressed the idea that even though the individual test may result in good scores for candidates it did not test their real communication skills. This argument is based on the possibility for candidates to develop exam strategies that could result in good grades while the learner may not really be competent. Secondly the individual test may not be valid in the eyes of participants because learners feel that their performance depends for a great part on the interviewer and his or her style. Some participants, (P 2 and P2) commented that emotional students might find the individual test particularly trying because they might be afraid of the interviewer. A smiling and supportive interviewer may inspire confidence and help candidates with their performances, while a different interviewer can make candidates loose all their means. The individual test is not the only test to have potential face validity issues; the paired test also contains element s that participants identified as potential sources of validity problems. In many cases participants believe that their performances could be different if they were paired with other partners or if they had to do the test alone. May participants reacted negatively to the presence of the partner whose level of language proficiency did not help them to complete their exam task successfully; or they reacted to the fact that the partner "stole" from them time for their performances or prevented them from remaining coherent and focused through their ideas.

In line with with studies which have found a positive correlation between face validity and learners' performances (Chan, Smidtt, & DeShon, 1996; James. Grand, Ryan, Schmitt & Hmurovic, 2010; Kajdasz, 2010) it was expected that the participants would work better on the

test that they preferred; because participants would choose as their preferred test the test which they found valid and which would consequently trigger a good performance. So, in order to measure the impact of such opinions on performances the study investigated the correlation between preferences and grades. The aim was to investigate whether participants' preferences influenced their performance. However, the present study did not reveal any significant relationship between those two variables; participants' performances seem independent from their preferences. The study did not did not find evidence of the impact of participants preferences over their performance. Can one therefore conclude the lack of importance of participants' perception on their score? The abundance of literature on the issue warns us against such a conclusion. It may be the case that due to the complexity of the context introduced by paired test and direct tests in general it is difficult to pin down score variance on one single element. One has to agree with Brown and McNamara (2004) that impact on score does not "support a simple deterministic idea" that an identified phenomenon "will have a direct and predictable impact on test processes and test outcomes"(p. 533).

However, according to their perception of their performance participants made choices which clearly indicated the test where they believed they would get higher scores even though those expectations did not match their actual grades. This mismatch between the actual scores and participants' expectations might come from the difficulties of scoring oral performances. But it mostly shows that participants assessed their production differently from examiners. Very often, as it appears in their answers during the interview, participants' assessment of their performance seemed to take only account of the quantity of language produced and the grammaticality of utterances, which is not the case for examiners and the rating scale they use. Since the purpose of the oral exam is to assess the proficiency of participants, examiners and the

rating scale takes into account most aspects involved in fluency. So this often results in a discrepancy between what participants expected and the actual grade they receive from examiners.

The lack of connection between grades and performances and the arguments provided by participants to justify their choices tend to support the idea that the participants did not have preconceived ideas about the test formats prior to the testing. Rather, their opinion regarding the test they preferred was formed according to the test conditions and the way they thought they performed. It is possible that a participant who opted for the individual test in the study arguing that he or she was nervous about the partner may have a different choice if that participant was paired with another partner. Likewise the one who expressed preference for individual testing because he or she liked the examiners' attitude, for example, might change his or her mind faced with a different examiner. For many participants what seemed essential was the conditions in which the test is delivered and how well those conditions helped them with their performances.

Limitations And Suggestions For Further Studies

There are a number of limitations in the current study. First, the study was conducted among ESL students; participants were informed that their participation and performances would not be reported to their class instructors. In other words the participants may not be the ideal population for such a study, because in terms of testing the context and the stakes play an important role test-takers' reactions. It is possible that the fact that the population was not sampled out of a real test population and there was no impact of the test on their lives or careers as there would be with a real oral proficiency test certainly could have affected learners'

reactions. The findings may be relevant in terms of examinees' reactions but one should be cautious about the generalization to actual ESL/EFL candidates.

Secondly, using a five bands rating scale for assessing participants performances during the study brought scores very close to one another. The differences between participants' grades in the paired test and the individual test were very small. Even the gaps between the scores of the more proficient speakers and those of the less able speakers were not always very large. This may have impacted analysis of correlations and other results. Future studies should consider adopting a 10 bands scale, which might reveal more spread between scores.

Another limitation of the study is linked to the design of the study; although the data collection instrument was the interview, at one point participants were asked to choose between two options; the paired test or the individual. The reactions of some participants suggested that there was a place for a middle or neutral position. Some students were undecided about the test they liked. Unfortunately the design of the study did not allow those participants to express this option. They were literally forced to choose between the two existing options; that is, they had to say whether they preferred the paired testing format or the individual format. With regard to the data I believe that a middle position would have appeared and probably shed a different light on the issue.

Finally there are some limitations related to the use of the interview as data collection instrument. Not all the participants had an appropriate level of language proficiency for an interview-based data collection. Since the study was using interview as data collection tool, it would be preferable if the participants were proficient enough to describe their feelings and express their opinions after the test, which was not always the case. This handicap, which future

studies could easily resolved, had certainly made us miss some worthy ideas that participants could communicate.

Despite these limitations this study has contributed to a better understanding of test-takers' preferences between individual and paired oral tests. The study has revealed which of the two tests most participants prefer and the reasons underlying their choices. Most previous research studies on learners' preferences have tended to justify and encourage the development of paired oral tests among candidates (Csepes, 2005; Együd & Glover, 2001; Saville & Hargreave, 1999; Taylor, 2001). However, candidates in those studies were not always given the option to choose between paired and individual test (Foot, 1999). Often, research methods did not include a comparison of participants' reactions and attitudes about the two test formats. The current study and the recent research study conducted by Marochi (2008) have adopted a different approach, which might explain why the two studies reached similar results. In both studies, participants took both tests consecutively and were interviewed about the tests. The advantage of such an approach is that participants were given the opportunity to compare and choose from two closely related test experiences. More studies based on such a direct experimentation with both test formats followed by a posttest interview could provide insight into learners perceptions about the two oral proficiency test format and which test they really prefer. Ideally such research should be conducted in exam settings; that is, participants should be sampled from an exam population and some limitations to the present study should be taken into account.

Implications

Given the limitations mentioned earlier, this study on learner preference is more nuanced than the previous research studies which have almost imposed the paired test as the most appropriate test format. Indeed, although the majority of participants preferred the individual test it seems premature to recommend immediate actions based on the findings.

A reason for such attitude is the fact that the study is one of the few existing studies on learner preference which appears as an important field of research in oral testing and assessment. Therefore more research studies seem necessary before one can fully understand learner preferences, their attitudes and their perceptions.

Moderation is also suggested by the results of the research; the findings of the study did not show any clear cause and effect relationship between preferences and performances. For example, where literature at large indicates a clear relationship between learners' perception of test validity and their performance (Brown, 2007; Chan, 1997), the study did not reveal evidence to support this claim. Further, one has to be careful with suggestions made by participants; many of them believed that they could have worked better and would have liked the paired test more if they were allowed a preparation time, as they were in the individual test. The preparation time awarded on the individual test seemed indeed, helpful for gathering ideas and making participants speak more fluently and develop their ideas more coherently. So they believed that in paired testing a similar period should also be awarded. However, such arrangement could change the nature of the paired test; we might go from natural and spontaneous speaking to a rehearsed dialogue, which some participants denounced as being a weakness of the individual test. So on this matter also, it seems that more studies on learners' preferences are needed. Such studies will have the advantage of testing the findings of the present study, and then contribute to

identify and enrich the database of phenomenon that are perceived by test-takers as potential sources of problems and score variance.

Finally, the results of this study suggest that both tests have their supporters even though the majority of participants preferred the individual test. Overall, few arguments provided by participants tend to show one of the tests as completely inappropriate for testing oral proficiency. Therefore more experimentations and reports are needed for stakeholders to have a better picture of both tests before any decision is made.

However, the answers of the participants revealed that candidates for ESL/EFL oral proficiency tests need some kind of training before the exam. The training could take the form of a special exam preparation or specific teacher attention during class. During classes, teachers should get students used to working in various modes; students should work individually, in pairs, and in groups, rather than limiting students to one way of working. Ideally the teachers should expose their future candidates to both exam formats so that candidates are not surprised by the exam format that they could meet. Equally, teachers and institutions preparing students for standardized tests should investigate the types of test their future candidates may have to take and help them with appropriate solutions. The oral exam training may also include helping ESL/EFL students to know how to assess their own oral production. It might be helpful for candidate to be familiar with the descriptors of an oral performance and how their performances would be scored.

Conclusion

Learner preference appears as a neglected research area as evidenced the number of publications on the issue. This area of research requires further exploration from a variety of perspectives. Such research will allow a better understanding of ESL/EFL learners' reactions and attitudes, and possibly lead to eliminating aspects of oral proficiency tests that might adversely affect performances and grades.

The present study was conducted through posttest (individual) interviews of 15 ESL/EFL students. It was an attempt to get insight into the preferences of the participants and the reasons behind their choices. The participants could communicate and comment on their opinions and feelings about the two oral test formats, since they had taken both tests and the delay between the tests and the interviews was very short.

Results indicate that the majority of the participants preferred the individual test to the paired test. This result is similar to the findings in Marochi (2008), but contradicts those published in other studies (Együd & Glover, 1999; Taylor, 2001). The reasons given by participants to support their choices relate first to the test items and some aspects of the test specification like the timing. Tests were generally preferred when the questions were brief, clear, and allowed participants to be fluent and coherent. From the participants' perspective these characteristics contributed to make the test appear easy. The reasons suggested by participants also relate to the impact of some anxiety factors. Factors such as the language level of the partner or the style of the examiner were viewed positively or negatively depending on the position taken by participants.

Although the study established clearly a preference of participants for the individual oral proficiency test, the results could not establish any connection between the participants' preferences and their performances. This result suggests the idea that participants did not have any set opinion about the tests, and may adapt to either test provided some factors are improved. Yet nothing really tells us if some the arguments that participants provided during the interview did not reflect some initial opinion they held about the tests. Therefore further research studies are needed are needed before one has a clearer picture of learner's preferences between individual and paired tests, the impact of those preferences on their performances, and the implications.

APPENDICES

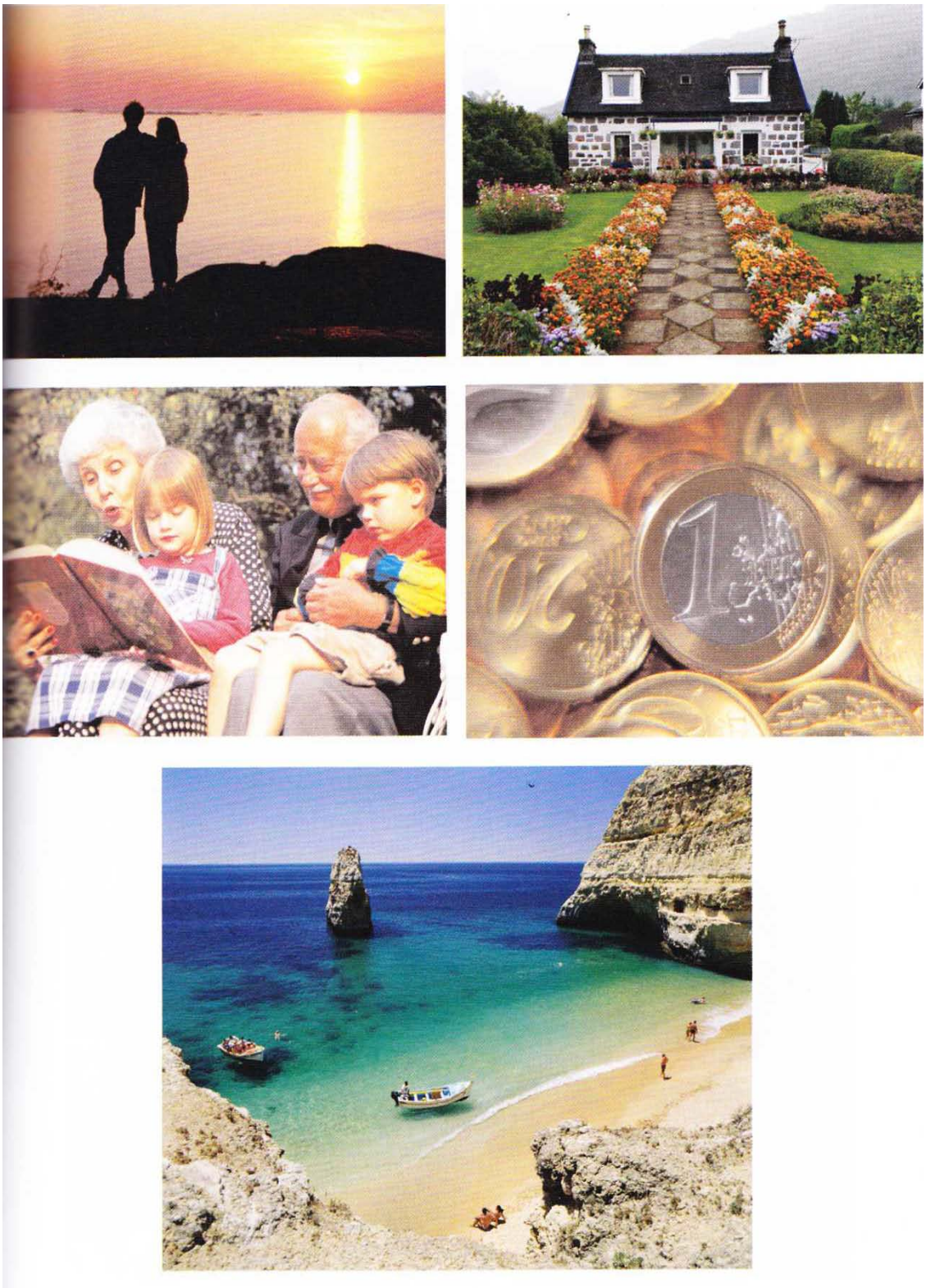
Figure 1: Pictures used in the paired test

Appendix A: Visual material for paired speaking test

copied from University Of Cambridge Local Examinations Syndicate (2008))

- How important are these things for a happy life?
- Which two are the most important?

Figure 1: (cont'd)



Appendix B: Paired testing material (interviewer frame)

Now, I'd like you to talk about something together for about three minutes

Here are some of the things in life which can offer happiness.

['Interlocutor' indicates the set of pictures to the candidate]

First, talk to each other about how important these things are for a happy life. Then decide which two are the most important. All right?

Examiner frame for Part 3, FCE speaking test.
Reprinted from: First Certificate in English for updated exam. (2008), P.105

Appendix C: Individual testing material

Long turn card

I would like you to tell me about the job you wanted to do when you were a child:

You should say:

- What that job was
- Why you liked that job
- If you still want to do it
- And say why or why not.

Examiner script

The Examiner will introduce part 2 by saying:

Examiner: Now, I'm going to give you a topic and I'd like you to talk about it for one to two minutes. Before you talk, you'll have one minute to think about what you're going to say. You can make some notes if you wish.

Do you understand?

Candidate: Yes, I do.

[Then the Examiner gives out some paper and a pencil for making notes, and read the topic.]

Examiner: I'd like you to tell me about the job you wanted to do when you were a child.

Adapted from Unit 8, IELTS Speaking test model, part 2: Recording script (CD 1, Track 22)

Appendix D: Semi-structured interview questions

Interview questions

1. You've just taken two oral tests; can you tell me how you feel about them? What are your impressions?
2. Which testing type allowed you to show your abilities?
3. What precisely helped you perform well on that test?
4. What would you like to change about the second test (on which you did not perform well)?
5. Did your class work prepare you for any of those tests?
6. Could you please describe your feelings during the tests?
7. In what way did your state --*interviewee's own word* -- influence your performance?
8. How did you like talking to a candidate like you rather than talking to the examiner?
9. For a future oral test which of these two test types would you rather have?
10. To finish I want you to describe me what your ideal oral test.

Appendix E: Oral paired testing exam rating scale

Adapted from Bonk, W. J. & Ockey, G (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89 – 110.

	Pronunciation	**Fluency**	**Grammar**	**Vocabulary/content**	**Communicative skills/strategies**
5	rarely mispronounces, able to speak with near native-like pronunciation	near native-like fluency, effortless, smooth, natural rhythm	uses high level discourse structures with near native-like accuracy	wide range of vocabulary with near native-like use, vocabulary is clearly appropriate to express opinion	confident and natural, asks others to expand on views, shows ability to negotiate meaning, shows how own and others' ideas are related
4.5 4.0	pronunciation is clear, occasionally mispronounces some words, but has mastered all sounds, accent may sound foreign but does not interfere with meaning	speaks with confidence, but has some unnatural pauses, some errors in speech rhythm, rarely gropes for words	range of grammatical structures but makes some errors, errors do not impede the meaning of the utterances	lexis sufficient for task although not always precisely used	generally confident, responds appropriately to others' opinions, shows ability to negotiate meaning
3.5 3.0	pronunciation is not native like but can be understood, mispronounces unfamiliar words, may not have mastered some sounds	speech is hesitant, some unnatural rephrasing and groping for words	relies mostly on simple (but generally accurate) sentences, has enough grammar to express meaning, complex sentences are used but often inaccurately	lexis generally adequate for expressing opinion but often used inaccurately	responds to others, shows agreement or disagreement to others' opinions
2.5 2.0	frequently mispronounces, accent often impedes meaning, difficult to understand even with concentrated listening	slow strained speech, constant groping for words and long unnatural pauses (except for routine phrases)	uses simple inaccurate sentences and fragmented phrases, doesn't have enough grammar to express opinions clearly	lexis not adequate for task, cannot express opinion	does not initiate interaction, produces monologue only, shows some turn taking, may say, 'I agree with you,' but does not relate ideas in explanation
1.5 1.0	Pronunciation frequently mispronounces, heavy accent, may use 0.5	fragments of speech that are so halting that conversation is virtually impossible	only says a few words, cannot make a reasonable judgment of student's grammatical ability	little lexis, inadequate for simple communication	may require prompting, shows no awareness of other speakers
00	does not discuss	does not discuss	does not discuss	does not discuss	does not discuss
Note: If a student shows she (or he) is consistently fulfilling the criterion, she receives the score at the top of the box, whereas if she only sometimes achieves the criterion level, she is given the lower score.					

P. Code ___ _____ _____ _____ _____ _____

P. Code ___ _____ _____ _____ _____ _____

P. Code ___ _____ _____ _____ _____ _____

Appendix F: Oral individual testing exam rating scale

Adapted from Bonk, W. J. & Ockey, G (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89 – 110.

	Pronunciation	Fluency	Grammar	Vocabulary/content
5	rarely mispronounces, able to speak with near native-like pronunciation	near native-like fluency, effortless, smooth, natural rhythm	uses high level discourse structures with near native-like accuracy	wide range of vocabulary with near native-like use, vocabulary is clearly appropriate to express opinion
4.5 4.0	pronunciation is clear, occasionally mispronounces some words, but has mastered all sounds, accent may sound foreign but does not interfere with meaning	speaks with confidence, but has some unnatural pauses, some errors in speech rhythm, rarely gropes for words	range of grammatical structures but makes some errors, errors do not impede the meaning of the utterances	lexis sufficient for task although not always precisely used
3.5 3.0	pronunciation is not native like but can be understood, mispronounces unfamiliar words, may not have mastered some sounds	speech is hesitant, some unnatural rephrasing and groping for words	relies mostly on simple (but generally accurate) sentences, has enough grammar to express meaning, complex sentences are used but often inaccurately	lexis generally adequate for expressing opinion but often used inaccurately
2.5 2.0	frequently mispronounces, accent often impedes meaning, difficult to understand even with concentrated listening	slow strained speech, constant groping for words and long unnatural pauses (except for routine phrases)	uses simple inaccurate sentences and fragmented phrases, doesn't have enough grammar to express opinions clearly	lexis not adequate for task, cannot express opinion
1.5 1.0	Pronunciation frequently mispronounces, heavy accent, may use 0.5	fragments of speech that are so halting that conversation is virtually impossible	only says a few words, cannot make a reasonable judgment of student's grammatical ability	little lexis, inadequate for simple communication
00	does not discuss	does not discuss	does not discuss	does not discuss
Note: If a student shows she (or he) is consistently fulfilling the criterion, she receives the score at the top of the box, whereas if she only sometimes achieves the criterion level, she is given the lower score.				

P. Code ____ _____ _____ _____ _____

P. Code ____ _____ _____ _____ _____

P. Code ____ _____ _____ _____ _____

P. Code ____ _____ _____ _____ _____

REFERENCES

REFERENCES

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching, 35,* 79–113.

American Council on the Teaching of Foreign Languages (1999). Retrieved from http://www.actfl.org/files/public/guidelinesspeak.pdf.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26,* 341-366.

Bonk, W. J., & Ockey, G (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20,* 89-110.

Brown, H. D. (2007). *Teaching by Principles: an interactive approach to Language pedagogy* 3rd Edition. New York, USA: Pearson Longman.

Brown, A., & McNamara, T. (2004). "The devil is in the detail": Researching gender issues in language assessment. *TESOL Quarterly, 38,* 524-538.

Csépes, I. (2005). Is testing speaking in pairs disadvantageous for students? A quantitative study of partner effects on oral test scores. novELTy , 9. Retrieved from http://www.lancs.ac.uk/fass/projects/examreform/Media/Article04.pdf.

Danesh, R. (2007). IELTS Speaking : Part 2 Topic Card. http://www.ielts-exam.net/index.php?option=com_content&task=section&id=6&Itemid=71.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26,* 367–396.

Douglass, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11,* 125-144.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26,* 423-443.

Együd, G., & Glover, P. (2001). Oral testing in pairs—a secondary school perspective. *ELT Journal, 55,* 70–76.

Fairclough, S. H., Tattersall, A. J., & Houston, K. (2006). Anxiety and performance in the British driving test. *Transportation Research Part F,* 9, 43–52

Foot, M. C. (1999). Relaxing in pairs. *ELT Journal, 53*, 36–41.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5*, 89–119.

Hong, E. (1999). Test anxiety, perceived test difficulty, and test performance: Temporal patterns of their effects. *Learning and Individual Differences, 11*, 431-455.

Horwitz, E. K. (1995). Student affective reactions and the teaching and learning of foreign languages. *Journal of Educational Research, 23, 569-652*.

Hsu, H. (2009). *The impact of implementing English proficiency tests as a graduation requirement at Taiwanese universities of technology*. (Doctoral dissertation, University of York). Retrieved from http://etheses.whiterose.ac.uk/576/1/PhD_thesis.pdf

IELTS. (2006). *Your IELTS Guide 2006-2009 http://www.yourieltsguide.com/ielts-test/ielts-score.html* Accessed 11/26/2011

In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System, 34*, 317–340

Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing, 5*, 51–66.

Grand, J. A., Ryan, A. M., Schmitt, N., & Hmurovic, J. (2010): How far does stereotype threat reach? The potential detriment of face validity in cognitive ability. *Testing, Human Performance, 24*, 1-28

Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment *The Modern Language Journal, 87*, 90-107.

Kajdasz, J. E. (2010). *Face validity and decision aid neglect.* (Doctoral dissertation). Available from ProQuest dissertations and theses database. (UMI No. 3438151)

Kivimäki, M. (1995). Test anxiety, below capacity performance, and poor test performance: Intrasubject approach with violin students. *Personality and Individual Differences, 18*, 47-55.

Lazaraton, A. (2006). Process and outcome in paired oral assessment. *ELT Journal 60, 287-289.*

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, N.J: Erlbaum.

Marochi, T. B. (2008). Individual and pair speaking test formats: A study of differences in performance. *Revista de Documentacao de Estudos em Linguistica Theorica e Aplicada (D.E.L.T.A),* 23-49.

Merriam, S. B. (2002). Introduction to qualitative research. In Sharan B. Merriam, (ed), *Qualitative research in practice* (pp. 1 – 33). San Francisco: Jossey-Bass.

May, L. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing, 1*, 29–51.

Norton, J. (2005). The paired format in the Cambridge speaking tests. *ELT Journal, 59*, 287–297.

Ohata, K. (2005). Language anxiety from the teacher's perspective: Interviews with seven experienced ESL/EFL teachers. *Journal of Language and Learning, 3,* 133-155.

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests.* Cambridge: UCLES/Cambridge University Press.

Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. *System, 30*, 143-154.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System, 28*, 373 - 386.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair task performance. *Language Testing, 19*, 277–295.

Porter, D. (1991). Affective factors in the assessment of oral interaction: gender and status. In S. Arnivan (Ed.), *Current developments in language testing* (92-102). Singapore: SEAMEO Regional Language Centre. Anthology Series *25*.

Richards, J. C. & Rodgers, T. S. (2001). *Approaches and methods in language teaching* (2nd Ed.). Cambridge: Cambridge University Press.

Saville, N., & Hargreaves, P. (1999). Assessing speaking in the revised FCE. *ELT Journal, 53*, 42-51.

Sawir E., (2005). Language difficulties of international students in Australia: The effects of prior learning experience. *International Education Journal, 6*, 567-580.

Taylor, L. (2001). The paired speaking test format: Recent studies. *Cambridge ESOL Research Notes, 6*, 15-17.

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation, 27*, 237-246.

University of Cambridge. (2008). *Reviewing FCE and CAE.* Retrieved from *http://www.cambridgeesol.org/assets/pdf/fcecae_review10.pdf.*

University Of Cambridge Local Examinations Syndicate (2008). *First Certificate in English 3 for updated exam*. Cambridge University Press, Cambridge, UK.

University of Cambridge. (2009). *Speaking test instructions: Cambridge ESOL Examinations. UCLES*. Retrieved from http://www.britishcouncil.org/217829_2009_speaking_test_instructions-022009.pdf .

Woods, P. (2006) *Qualitative Research*, retrieved from http://www.edu.plymouth.ac.uk/resined/Qualitative%20methods%202/qualrshm.htm.