INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Self-Assessment of Communicative Ability: Investigation of a Novel Tool for

ESL Learners

Edith Julie Dandenault

A Thesis

in

The TESL Centre

Presented in Partial Fulfilment

of the Requirements for the

Degree of Master of Arts at

Concordia University

Montreal, Quebec, Canada

May 1997

Canada

ABSTRACT

Self-Assessment of Communicative Ability: Investigation of a novel tool for

ESL Learners

Edith Julie Dandenault

This thesis reports an investigation into the development and use of a self-assessment tool in the field of ESL.

The study involved a group of Francophones telling, in their own words, the story represented in a wordless cartoon storyboard. Segments lasting 30 seconds were taken from each of the recorded story tellings and played to a large group of native speakers of English. These English speakers rated the voices on a seven point scale where 7 = "can express ideas in English like a native speaker" and 1 = "cannot express ideas in English at all". From the ratings obtained we found voice samples that were statistically distanced from each other so that they lay on a seven point scale at approximately the 2,4 and 6 points of the scale. A new stimulus tape was made up using two samples from the 2, 4 and 6 categories of speaker, as well as one speaker in the "2" category and one in the "7" category. This tape comprised the stimulus tape for the final part of the study. A group of French speaking students who were learning English as a second language rated themselves against these selected voices by indicating for each whether they could express themselves better than, as well as or worse than each sample voice.

The results were submitted to the Rasch model of measurement (item

response theory) which demonstrated that the learners were extremely

consistent in the way they rated themselves. Overall, the study indicated that

this test, which is relatively simple to administer once the stimulus tape has

been made, has strong face validity and strong internal consistency. The

potential implications of these results for use of the tool in self-assessment as

a placement test are also considered.

ACKNOWLEDGMENTS

TABLE OF CONTENTS

List of Tables

CHAPTER I: INTRODUCTION

As defined by Bachman (1990), evaluation is the "systematic gathering of information for the purpose of making decisions" (p.22). The key expression in this albeit general definition is *making decisions*. In English as a Second Language (ESL) contexts, decisions are constantly being made about students and their learning. Examples of such decisions are: (1) insuring that students are placed in the right starting class, (2) determining how much progress was made over a fixed period of time or (3) determining whether students should be promoted to more advanced classes. All these decisions are taken on the basis of the results of one form or another of evaluation. For instance, if a teacher wants to determine if a student will be promoted to the next level within an ESL program, the use of an achievement test designed within that program will most likely be administered as opposed to the TOEFL test which better serves as a general proficiency test. So, in fact, the decision which needs to be made dictates the type of test required.

The following study is a report on the creation and administration of a specific type of test (self-assessment) designed to be used in the context of placement of adult learners in the in ESL classroom.

Testing in Second Language Acquisition (SLA) Research.

Reporting on the state of testing research in Second Language Acquisition (SLA), Bachman (1990) claimed that investigators have, for the

most part, concentrated their efforts on reworking existing tools or re-evaluating currently used materials rather than thinking up new ways of evaluating SLA. While updating existing tools is necessary, he suggests that new and more efficient means of testing should be found. He warns that lack of new ideas and new developments in testing may in fact lead to an undesirable standstill in testing research (see also Anastasi, 1989). In order to make appreciable contributions, however, evaluation tools need to be designed with a very specific context in mind, since the creation of testing instruments necessarily starts with the identification of its intended primary use (Bachman, 1990).

Examples of the types of testing used in the field of ESL include the following: (1) proficiency testing which involves determining a person's general ability in a language or their ability with a specific skill; (2) diagnostic testing, which is designed to signal the strengths and weaknesses of the student in the language; (3) progress reports which are used during the course of a language program to indicate whether students are in fact meeting the pre-determined objectives; (4) achievement tests, which are usually administered after a specific teaching/learning session to reveal which final objectives of the course were attained; or (5) placement tests, which are used prior to the entry into a particular course to place students in the appropriate level. Of these test types, one of the less frequently discussed is placement testing.

Placement testing: Context for development

The most common type of placement test in current use is the oral interview. Teachers generally find, however, that oral interviews are difficult to use. Not only are they time consuming to administer, but to minimize subjectivity they often require the services of at least two people. Oral interviews therefore do not meet the requirements of economy and ease of administration suggested by Oscarson (1980):

> [placement tests] allow the individual to enter a learning system at the appropriate point and to give indications of the kind of learning system appropriate to him. The most important criteria for such tests are *economy* and *ease of administration* in relation to the reliability of the assessment and perhaps prognosis provided (p. viii, italics mine).

There is, therefore, clearly a need to find a placement tool which eliminates such problems. Self-assessment is a promising possibility as it can be made both to be easy to administer and to be time-saving. Note here that the term "self-assessment" is being used instead of others in the literature such as "self-rating", "self-evaluation" or "self-appraisal". The term self-assessment is a less*loaded* term in that it does not carry such a final and evaluative connotation (Oscarson, 1980) and as well is favoured in the literature (Blanche & Merino, 1989).

Self-Assessment in Testing

A number of authors have tried to encourage researchers to pay more attention to self-assessment and promote its use. In a discussion on the place of innovation in testing, Alderson (1986) suggested that the recent push by teachers to involve the learner in the decision-making process is one reason for using self-assessment. Involving the students in the evaluation process of their learning is believed to help activate their monitor, which, in turn, allows them to know where they stand with regard to program objectives (Holec, 1981). Self-assessment also helps to increase the autonomy of the learner (Oscarson, 1980). Wangsotorn (1980) speculates that this new gained autonomy increases the ability of learners to make autonomous judgements which in turn helps them better direct their foreign language learning. The use of self-assessment in the classroom also ensures that the burden of evaluation is shared by both the student and the teacher (Oscarson, 1989). Spreading the burden of the evaluation process means mutual responsibility for learning. Because learners and teachers often see things differently and often establish different priorities, having them share the responsibility is likely to encourage a change of emphasis in classes. Finally, studying self-assessment as a means of evaluation is an important educational objective in its own right (Oscarson, 1980). We can suppose that a reason for this is that it develops the ability to evaluate the effectiveness of one's own performance which is part of learning how to learn, or that it may help learners to better

4

set their own goals and objectives throughout the learning process.

Of course, no matter how convincing the arguments for using self-assessment may be, Dickinson (1987) suggests it will not be easy for teachers to accept self-assessment as a legitimate means of evaluation. He feels that teachers will always believe themselves to be more able to make accurate assessments of their students' abilities than the students themselves. The novelty of self-assessment as a placement measure also does not help make it a convincing alternative to what teachers are used to.

Self-assessment tools in the past have most often involved placing students in a specific hypothetical linguistic context and asking them to assess their own ability in this given situation. The standard tool used is the questionnaire, but the benefits claimed for the use of this type of self-assessment are not uniformly supported by the findings.

Based on her study of 22 adult learners of ESL, Anderson (1982) found that students' self-perceptions did not match teachers' perception of these same students and that students from certain cultures tended to underrate their abilities. In his research conducted with adult learners of Swedish completing self-assessment questionnaires, von Elek (1982) found a high correlation between teacher ratings and self-assessment results (see also Wangsotorn 1980), but like Anderson, he found that students tended to underrate their abilities. In 1978, Ferguson looked at the results of self-assessments on three different types of questionnaires and came to the

conclusion that, in fact, students in general tended to overestimate their abilities. Note here that in both of these studies the ability of the students were measured by the scores obtained on standardized ability tests. Finally, based on the findings of ongoing research on placement testing at University of Ottawa, Leblanc and Painchaud (1985) reported that students can self-assess well when it comes to their abilities in their second language. As one can see, the conflicting findings on the issues of correlations with teacher ratings and overestimation alone indicate the difficulty involved in reaching any conclusions about the actual value of self-assessment. Such differences in findings could be due, in part, to some of the following factors.

First, the reality behind the use of self-assessment is that it places a certain burden on the learner. Most tests require students to rely on memory of their abilities in order to rate themselves. Is it realistic to expect them to do so efficiently? This question forces one to take into account learner abilities such as memory, ability to recall information, or other related skills, which are, in fact, separate from the actual ability to self-assess.

A second aspect of self-assessment which may strain the learner is the fact that most self-assessment tests focus on grammar, vocabulary, and syntax, but not communication. Today, most students learn in order to gain skills in communication. Given this motivation, it is likely that students will expect the self-assessment tests administered to reflect this aspect of learning. It is possible that students will have difficulty with self-assessment tests which do

not measure what they expect it to measure.

Finally, one might want to know what each student's point of reference is when he or she thinks about his or her own ability in their second language. How far back in time is the standard to which they are comparing themselves? Would having a set of samples with which to compare themselves, in fact, help them better determine their own ability?

The new self-assessment tool designed and investigated in this study tried to address some of these challenges. It involved students self-assessing their abilities in their second language by rating themselves against recordings of carefully chosen voices of speakers predetermined to be at different levels of proficiency in English. It was hypothesized that this tool would remedy some of the ailments which have been associated with self-assessment by (1) having participants imagine accomplishing one very specific task as opposed to requiring them to think about their general ability in their second language, (2) involving participants in a communication based task (telling a story), and (3) providing participants with specific points of reference against which to compare themselves and thereby eliminating the need for them to rely on memory of past communicative experiences.

CHAPTER 2 : REVIEW OF THE LITERATURE

The purpose of the present study is to investigate the construction and use of a self-assessment tool which requires participants to rate themselves against pre-recorded samples of speech. This study was prompted by the desire to have a practical tool for self-assessment that is easy to administer. The following section presents a review of the literature which establishes the context for such an investigation.

In the field of applied linguistics, testing is considered by some to be a peripheral issue, of interest only to those who engage in it (Bachman, 1991). Bachman suggests that such a separation is unproductive, considering that testing and evaluation are both involved in second language learning (see also Douglas, 1975. Whether it comes at the beginning of a program in the shape of a placement test, at the middle as a progress report or at the end as an achievement test, evaluation is very much part of the second language learning process. Despite this apparent separation of testing from the more general issues in SLA research, a literature review reveals attempts made by applied linguists interested in evaluation to integrate language testing research. The direction future studies should take to favour this integration of testing has only recently been the object of discussion.

Although Bachman (1991) claimed that in the past ten years the testing research has focused on the re-evaluation of currently-used tools, the development of more communicative tests, and the changing of theoretical

views in the field, he suggested that discussion of directions in testing has become the focus of the last decade only . Douglas (1995), in a state- of-the-art paper, summarized the course of such developments to conclude that on the whole the research has been rather fragmented. He does, however, also reiterate the fact that the entire field of SLA could benefit from advances in testing because of its close link to second language learning. One particular area of SLA which could benefit from the research conducted in evaluation is test construction.

As early as 1986, Alderson urged researchers to attempt to create kinds of tests that are inspired by a new approach in which testing is part of the learning process. Coining the expression "tests as innovators", Alderson holds that the influence tests have on learning, be they positive, negative or neutral, should act as justification and motivation for creating tests that are closely linked to teaching.

More recently, another call for the investigation into test construction was made by Hamayan (1995). She claims that the nature of language testing must change because students' expectations and objectives for learning ESL have changed over the past two decades. Students are now turning toward language classes for more communicative purposes. Hamayan, however, does not discuss how new tests should accommodate this shift to a communicative focus.

One clear conclusion which can be reached by looking at the above

statements is that investigation into new approaches is being encouraged. In order then to begin investigation into creating new tools, the literature which discusses the different aspects and requirements involved in creating them needs to be presented.

Some of the more widely discussed and studied concepts involved in test creation have been the following: (1) effects of testing on learners (Berkoff, 1985, Whiteson, 1981), (2) distinction between performance and competence (Bachman, 1990, 1991), (3) backwash effects in testing (Promoudou, 1995, Whiteson, 1981), (4) integrative vs discrete point testing (Farhady, 1979, Laesch & von Kleek, 1987), (5) importance of goal setting in test preparation (Bachman, 1990), (6) implications of test administration (Clark & Clifford, 1988), to mention only a few. Although all of these issues are intimately linked to the creation of new tools, the context within which a test is designed remains the most important determinant of the type of test which will be required. In the past, various contexts have been investigated in SLA testing research.

As described in Chapter 1, some of the more common tests used are: diagnostic, proficiency, achievement, progress, and placement testing. Of these, proficiency testing has received a large share of researcher's attention, because of its obvious importance to the field of SLA. Teachers are also possibly more familiar with such tests because they are often involved in creating and modifying them. In contrast, very little has been written about

placement tests. Despite the fact that they are used frequently, placement tests are seldom prepared by teachers, so they place little interest in their conception (Wall, Clapham & Alderson, 1994). Furthermore, because finding a placement test that meets the specific needs of every teaching/learning context is impossible, each placement test is designed for a particular situation. For this reason, placement tests often have limited generalizability. Of course, there are some commonly used placement testing procedures such as cloze (Aitken, 1977, Bachman, 1985, Hale, Stansfield, Rock, Hicks, Butler & Oller, 1989), oral interviews (Bachman & Savignon 1986, Burt & Dulay, 1978) and various forms of traditional objective questionnaires (see Rossi, 1983 for complete review). However, unless a test is as well-known as the TOEFL, there is virtually no way it could be understood or appreciated by any institution other than the one for which it was created (Wall et al. ,1994).

It is reasonable to assume that these issues are the cause for less time and effort being invested in the development of new placement tools. Investigations into new avenues for placement might not only be interesting, but also practical and necessary as current tests are generally found to be too time consuming (Mullen, 1978) and their administration a misuse of teacher time (Dickinson, 1987). One particular type of placement tool which has been the focus of recent studies as a possible remedy to the above described problems is self-assessment.

Unlike teachers who may have reason for not trusting self-assessment,

many researchers recognize self-assessment as an effective, trustworthy placement tool. There are several reasons for this. Some researchers argue that the learner is the one who knows him or herself best (Raasch, 1978). Others have claimed that most learners can give good estimates of their ability in a second language (Upshur, 1975). Finally, some have presented the argument that given the proper instruments adults are capable of self-assessing (von Elek, 1982).

The Council for Cultural Cooperation of the Council of Europe, in 1971 (cited in Bachman, 1990), brought up self-assessment as an interesting tool for needs analyses of learners. Prior to that, self-assessment was used mostly in the fields of psychology and social sciences, not as a tool to measure language ability, but as an instrument to study subjects' perceptions.

Almost 10 years later, Oscarson (1980) wrote "Broadly speaking, self-assessment is a new field in language testing and consequently there is very little accumulated knowledge and experience to draw on (...)". (p. 13), indicating that despite the interest signalled by the Council of Europe little research was, in fact, conducted. To this day, although somewhat larger, the size of this accumulated knowledge still remains relatively small. This could be due to the fact that the first tools designed for self-assessment were merely self-administration and self-correction of regular tests. It was not until recently that tools were designed especially for self-assessment of one's abilities (Oscarson, 1980). Since the mid-80s, however, there have been some

very interesting developments in the research.

Some have investigated the reliability of self-assessment tools and found that little trust could be placed on certain types of self-assessment tools (see Davidson & Henning, 1985), while others found self-assessment to be a relatively good evaluation tool (see Bachman & Palmer, 1988, Janssen-van Dieten, 1989, Leblanc & Painchaud, 1985). Others examined the ability of learners to rate themselves and found that some overrated their general abilities in the second language (Heilenman, 1990; Janssen-van Dieten, 1989 while others underrated it (Davidson & Henning, 1985; Oscarson, 1980).

The above investigations and others like them underscore the interest people have in different aspects of self-assessment. Oscarson (1989) reports six reasons for wanting to look into self-assessment. First, self-assessment may promote learning as it encourages autonomous judgements of the effectiveness of one's communication (see also Lewkewitz & Moon, 1985, Wangsotorn, 1980). Second, by encouraging a raised level of awareness, self-assessment stimulates learners to consider course content and fosters evaluative attitudes in the learner. Third, improved goal orientation which in encouraged by self-assessment should enhance the learner's knowledge of possible goals helping exert control over their own learning. Fourth, self-assessment may help broaden the learner's perspectives of evaluation of communicative competence. Fifth, self-assessment may alleviate the assessment burden on the teacher, allowing him or her to be freed for other

important duties. Finally, teaching students how to continue learning the language autonomously after a course is universally considered as an important objective in foreign language instruction. Such reasons for promoting self-assessment in general are, of course, essential to its growth, but there are also reasons for wanting to use self-assessment specifically in the context of placement. Quantitative research such as that reported in Blanche (1990), Ferguson (1978), Raasch (1978), and von Elek (1982) provides a few such reasons.

In 1978, Raasch reported using two questionnaires designed to work together to determine students' progress in their second language classrooms. One questionnaire was for teachers to use in assessing students and the second one was for students to assess themselves. The test items required students to place themselves in a specific context and determine how well they would use the target language in that context. Raasch supports the idea of using self-assessment in the classroom as he maintains that it helps focus the learner's efforts and provides both the teacher and learner with valuable information on student achievement of goals. He also suggests that past research has shown the precious information self-assessment provides teachers. Unfortunately, he does not present any statistical support for these statements.

Also in 1978, Ferguson proposed his own set of three self-assessment tests. The first of the three tools involved the self-assessment of 35 speaking

skills. Students answered questions like: "Can you order a simple meal?" by yes or no. The second tool involved students comparing themselves with 18 taped voices. Students heard 18 English voices and had to choose the first voice that was better than them. Finally, the third tool involved the speed of identification of syntactic errors. The three tests were administered to 89 students at the University of Geneva and 90 students at an interpreter's school. The subjects' abilities and experience in English varied from those who had only 50 hours of study and exposure to English, to native speakers. The correlations between the results of these three tools and standardized measurements of fluency and syntactic correctness of spoken English were respectively r=0.39, r=0.19, r=0.87 (no probabilities presented). These results led the author to conclude that the third test, that is the one involving the speed of error recognition, might be a very useful and valid self-assessment tool, which could be used as a placement test.

Later, in a study of adult migrants studying Swedish as a second language, Von Elek (1982) developed his own complex and extensive series of self-assessment questions as part of a diagnostically oriented proficiency test of Swedish. His questionnaire was a six skill matrix (vocabulary, grammar, listening comprehension, reading comprehension, oral production, and written production) with ten levels of ability for each skill. When a student could answer 80% of the questions from each skill at one level, he or she could move on to the next level. The student stopped changing levels

once he or she could no longer answer 80% of the questions. The test results were compared with teacher ratings and were found to correlate only moderately (no specific correlation figures were reported).

Finally, Blanche (1990) prepared a self-assessment scale for 43 American students of French enrolled in the DLIFLC (Department of Defense Language Institute/Foreign Language Center) six month French language course. In the study students participated in a role-play with their instructor after which they were asked to say what grade they thought they would receive for their performance. They were given a scale on which to place themselves. The overall result suggested that there was no significant relationship between the participants' self-appraisal and their graded performance.

The findings of the above mentioned studies do not allow the reader to draw clear conclusions about the value of self-assessment because there does not seem to be a consensus. This inconclusiveness is reflected further in the works of other authors such as Anderson (1982), Hale et al. (1989), Janssen-van Dieten (1989), Oscarson (1980), Wangsotorn (1980), and Wesche, Morrison, Ready & Pawley (1990), who found anywhere between no correlations and high correlations between self-assessment and standard more objective evaluation tools. Such rather inconclusive results may be the best available, so they indicate that continuing investigation is required. In an attempt to help guide further investigations, Oscarson (1989) claimed that efficient self-assessment would need to involve the learners rating

themselves against a clearly defined scale (yardstick). This point was also made by Leblanc & Painchaud (1985) in the context of their ongoing research on the use of self-assessment in placement testing. The exploration of such a self-assessment tool is precisely the aim of this thesis.

In a study on communication in non-fluent bilinguals, Segalowitz (1976) used a specific type of self-assessment tool to screen participants. Segalowitz described the self-assessment test as follows:

> The assessment consisted of a voice comparison task in which the subjects compared their ability with that of three prerecorded Anglophones to retell a simple story in French. These prerecorded samples had already been judged by Francophones to lie at points 2,4,6 with non-overlapping distributions in the ratings on a seven point scale ranging from cannot express any ideas in French at all to can express ideas in French as well as a native speaker. Subjects indicated whether they believed they would do worse than, about the same as, or better than each of the taped Anglophones. (1976, p. 124)

After self-assessing against the voices, the same subjects were recorded while telling the story and then judged by native speakers of French. The results of the comparison between the self-assessment scores and those given by the native speakers inspired confidence in the test since the way people rated themselves was actually close to how native speakers rated them

(Segalowitz, personal communication, March 1997). Furthermore, according to the author, this self-assessment method better predicted the way native speakers of French judged the subjects' speech samples than did the self-rating scores obtained with a simple 7-point Likert-type scale, a method advocated by Macnamara (1967). Oscarson (1978) might have predicted that this would be the case as he maintained that provided with a standard with which to compare themselves learners could determine their language ability. Unfortunately, however, the details of validity and reliability of this new tool were not reported.

The present study involved creating a type of self-assessment tool similar to the one used in Segalowitz (1976). This tool (Voice Comparison Test or VCT hereafter) was administered to Francophone second language learners of English who normally write a standardized placement test upon entry into the program. After the administration of the tests, the following four separate aspects of the instrument were investigated: (1) the internal consistency of the self-assessment tool; (2) the evidence for validity of the scale used in the self-assessment tool; and (3) the relationship between a person's ability to make consistent self-assessments and individual characteristics such as age, sex, motivation (Anderson, 1982), attitude (Blanche, 1990), age of first contact with second language, and level of ability in the second language, and (4) the relationship between the results generated by this self-assessment tool and those generated by placement tests in current

18

use.

The next chapter describes the details and procedures involved in the development of the Voice Comparison Test. Before taking up this discussion, however, it would be useful to clarify the logic of the research and the analysis of the data.

The self-assessment ratings were submitted for analysis to the Rasch Model of Measurement (Item Response Theory–IRT) (Linacre, 1989-1993; Wright & Stone, 1979). This analysis generates a model of the data that includes a scale along which the stimulus voices are located and along which the individual ESL learners are located. The responses to the various items are statistically adjusted through a series of iterative analyses so as to achieve the best fit possible for the stimulus voices, based on subjects' self-assessment ratings against these voices. As indicated below, in the process of generating the best fitting model, the Rasch analysis provides an number of indices that proved to be useful for the research questions of interest in this investigation.

For the Rasch analysis to be applicable, the data must meet a certain set of requirements. First, the data (here, the self-assessment ratings) must be reduced to a unidimensional abstraction (here, language proficiency) and all test items must be locally independent of each other. Second, the data must involve a more-or-less comparison (here, this was in terms of judging one's own ability to speak to be better than, as good as, or worse than the sample voice). Third, the data must involve a linear scale, that is, items must be

locatable along an abstract linear continuum. In this study, the computational

procedures placed both self-assessed speaking ability and difficulty of the

items (ability level represented by a stimulus voice) on the same scale.

Finally, it must be possible to determine a unit of measurement by a process

that is replicable without modification over the entire range of the variable.

In the present case this accomplished by the Rasch analysis in terms of logit

units. The data obtained from the VCT met all of the above criteria, and thus

it was possible to apply the Rasch Model of analysis.

The first aspect of the investigation concerned the reliability of the tool.

It was not feasible to test subjects more than once to obtain a test-retest

reliability measure. The Rasch Model of Measurement could, however, be

used to provide a measure of the internal consistency of subjects' responses.

Essentially, the Rasch Model provided an index of the degree to which

subjects' rating data fit the model generated by the analysis. Departures from a

good fit would reflect inconsistencies in responding. Such departures will

arise when subjects do not treat stimuli in a consistent and coherent manner.

The data may be considered reliable to the extent that subjects rate the stimuli

in a manner that is consistent with an overall model of the data–that is, to

the extent that there is little noise in the data and the overall fit is strong.

The second aspect of the investigation concerned the validity of the

tool. Testing the validity of an instrument is always problematic since it is

never easy to know what external criteria to invoke in order to assess validity.

In the present research, there was a phase in the design of the VCT in which the voices were selected for the final version of the tape. This selection was based on the ratings given by native speakers of English about how well each speaker of the taped samples could express themselves in English (ranging from "not at all" to "like a native speaker"). These ratings placed the voice samples at points 2, 4 and 6 on a 7 point scale. It was assumed that this ranking of the voice samples by native speakers had face validity as indications about the true nature for the voice samples. The Rasch analysis of the ratings by the 135 ESL subjects (total of 1112 responses) generated a scale upon which the voice samples were distributed. The extent to which the locations of the voice samples on this scale, based on the ESL learners' ratings, matched the ordering of the voice samples by the native English speakers provided an index of the validity of the test, that is, of how well the test produced results that corresponded to reality as indicated by the native speakers judgments.

A third aspect of the investigation involved the relationship between ability to make consistent self-assessments and certain characteristics. Here we asked the question of whether characteristics such as age, sex, attitude and motivation, and length of study predicted ability to self-assess consistently. The ability to self-assess consistently was indexed in this research by a measure provided by the Rasch analysis called the "infit coefficient". The infit coefficient indicates to what extent a given individual provided data that

were consistent with the way the overall model related self-assessment ratings to the placement of stimulus voices along the scale. Some subjects may be expected, for example, to provide data which, in comparison to the data provided by other subjects, are relatively more noisy or do not fit the general pattern very well. It becomes possible, therefore, to see whether this fit of individual subjects varies significantly between groups (males/females; older/younger; etc.). In this way it becomes possible to test the relationship between individual difference factors and the consistency with which one self-assesses.

The final aspect of the investigation concerned how the placement of the individual ESL learners on the scale generated by the Rasch analysis of the self-assessment ratings would relate to the placement of the same learners on the scale given by another assessment procedure which is commonly used in the school system.

CHAPTER 3 : RESEARCH DESIGN

This chapter presents the various steps involved in developing and testing the self-assessment tool being described here. Briefly, these included the following: (1) collecting the initial, large set of voice samples in French and English for each speaker; (2) submitting the <u>French</u> language samples to a screening by native speakers of French in order to eliminate potentially problematic voice samples; (3) editing the <u>English</u> voice samples of the speakers whose voice samples were retained in Phase 2 in order to produce a stimulus tape; (4) submitting the stimulus tape to native speakers of English for judgments; (5) selecting the final set of English voice samples based on the judgements made in the previous phase and constructing the self-assessment tool with these voice samples; (6) designing the questionnaire to be given to users of the self assessment tool, and (7) administering the VCT to a sample of French speakers learning English as a second language. Each of these seven phases involved a number of detailed research design considerations that are discussed later in this chapter. For the reader's convenience, Table 3.1 summarizes the principal details in each of these phases by showing which participants were involved in a given phase, what materials were used, and what procedures were followed.

The data obtained in Phase 7–involving the administration of the self-

Table 3.1.

<u>**Summary of Phases Involved in Study**</u>

Phase	Name	Participants	Material	Task/Procedure	Outcome
1	Initial voices	- 22 native speakers of French. - Males aged between 19-36	– Storyboard	- Participants told a story in French and then in English based on a storyboard. - Participants were recorded	- 22 French versions of the story and 22 English versions of the story, each "telling" lasted 1-3 minutes.
2	Initial screening of voices	- 10 native speakers of French	- Bi-polar personality scales - French recordings from PHASE 1	- Native speakers listened to the French recordings and rated them on the bi-polar personality scales	- Selection of 16 speakers whose French voices did not elicit extreme perceptions of personality traits
3	Editing		- English samples of speech of the speakers selected in PHASE 2.	- Samples were digitized on a Macintosh Quadra 840 AV and edited down to 30 sec. segments	- 2 tapes with different orders of the 16 English selected speech samples (Tape I and Tape II)
4	Judge-ment	- 29 native speakers of English - Mix of ESL teachers (6) and workers from unrelated fields	- Tapes created in PHASE 3 - Four 7-point scales: accent, ease of speech, ability to express ideas, overall ability	- 14 judges rated the 16 voices from Tape I - 15 judges rated the 16 voices from Tape II	- Statistics for all 16 voices
5	Selection for final tape		- Statistics from PHASE 4	- Calculated mean scores for each voice t-tests conducted to identify statistically separated voices	- Selection of 8 voices and creation of the final Tape
6	Ques-tionnaire design		- Background & consent - Storyboard - Self-assessment scales - Motivation and attitude questionnaires	- Design and assembly of each section of the questionnaire	- Final questionnaire
7	Test adminis-tration	- 149 students - CEGEP Sherbrooke - Ages : 16-43 - Groups of 25-34	- Stimulus tape from PHASE 5 - Questionnaire from PHASE 6	- Test administration	- Data ready for analysis

24

assessment tool to a group of language learners–were submitted to various analyses that addressed the following four questions concerning the validity, reliability and other characteristics of the tool:

First, does this self-assessment tool meet an appropriate standard of reliability as reflected in the internal consistency with which students make their ratings?

Second, is there evidence for the validity of this tool as a measure of speaking ability?

Third, do the results provided by the tool offer interesting insights into the relationship, if any, between language learners' perception of their own abilities and individual characteristics such as age, sex, motivation, attitude, level of ability, or age at which one starts to learn the language?

Finally, how do the results yielded by this tool compare with those obtained with other tools in current use?

We now turn to a more detailed description of the procedures carried out for each of the seven phases.

Phase 1: Initial set of voices

The purpose of this phase was to develop a large set of potential stimulus voices from which a smaller set could be chosen for use in the self-assessment tool.

Participants

The participants in this phase of the study were 22 male native speakers of French (Québécois) between 19 and 36 years of age. All participants signed a consent form before taking part (see Appendix A).

Materials

The material required was a storyboard which consisted of a cartoon strip of 12 wordless pictures depicting a matrimonial dispute (Appendix B).

Procedure

Each participant was recorded telling the story depicted in the storyboard. This task was chosen to provide speech samples because in significant ways it recreates many of the demands speakers encounter in the real world. Here participants had to describe a situation they were viewing, just as in real life one has to recount an event or describe something which just took place (Clark & Lett, 1988). Prior to being recorded in both French and English, the participants were given time to look at the pictures and encouraged to ask questions. They were asked to tell the story as they would if they were speaking to a friend and were asked to talk for at least 2 minutes. All participants first told the story in their mother tongue then in their second language and were aware that they were being recorded.

Results

This phase of the study yielded French and English speech samples of 22 speakers from which a critical set of samples could be selected for use in the assessment tool.

Phase 2: Initial Screening of voices

The self-assessment tool being developed here was to involve learners listening to voice samples and making judgments about their own skill level in comparison to the samples they were listening to. In this situation, one always has to worry about the possibility that some characteristic of the speakers' voice might systematically influence the listener even though that characteristic is, from a logical point of view, irrelevant to the judgment being made. For example, a voice sample may be interpreted as conveying anger or sarcasm or reflecting some personality trait of the speaker. This may influence the listener's judgment (e.g., the listener may judge him or herself to be different from the speaker because of a desire to distance himself or herself from the perceived trait of the speaker. The purpose of this phase of the research, therefore, was to identify voice samples that appeared to elicit systematic trait perceptions from listeners. Once identified, these voice samples could then be excluded from the final tool.

Participants

Ten adult native speakers of French rated the voices on a series of bi-polar personality scales designed to help screen for personality features which might be systematically reflected by the voices in the speech samples. The raters were students and professionals having no particular training in the areas of teaching second languages.

Materials

Thirty second samples of the 29 voices recorded in Phase 1 were recorded onto a stimulus tape which was played in full to all ten Francophone raters.

Raters were given a personality scale adapted from Ryan and Caranza (1975) which consisted of the following items in French: instruite/pas instruite, riche/pas riche, accomplie/pas accomplie, intelligente/pas intelligente, fiable/pas fiable, gentille/pas gentille, sympathique/pas sympathique, généreuse/pas généreuse, honnête/pas honnête, travaillante/pas travaillante (see Appendix C). The raters were asked to evaluate each of the voices on each of these characteristics on a scale of 1-7 (7 indicating that the voice heard was characterised by that feature).

Procedure

The rating sessions took place in three different locations with small

groups (3 - 4) or raters. Each rating session was led by the researcher in order to ensure that the same directions were given to all raters. The tape with the 29 recordings was played only once. The raters were asked to judge each of the voices on the different bi-polar personality scales as they listened to the recordings. There was a 10 second pause recorded onto the tape between each voice.

Results

Based on the judgements made by the 10 francophones, voices which rated at the lesser (1) or greater (7) end of any of the bi-polar scales were eliminated from the voice samples which were to be used in the next phase.

Phase 3 : Editing the retained voice samples

In Phase 2, 6 French voice samples identified as eliciting inappropriately strong perceptions of trait characteristics were eliminated from the sample group. The English speech samples corresponding to the retained voices were thus eligible for this next phase of the tool development. These 16 English voice samples were digitized onto a Macintosh Quadra 840 AV to allow editing down to 30 second segments. In all cases, both the very beginning and the very end of any story were eliminated. This was done in order to eliminate false starts and disorganized endings in any voice sample. This editing process also allowed for the elimination of particularly long

pauses.

To ensure that no particular voice was disadvantaged by its position on the tape in the judging process, two versions of the tape were prepared, one version being exactly in the reverse order of the other.

Phase 4 : Judgements by native speakers of English

The purpose of this phase was to obtain an initial scaling for the voice samples which would be used in the Voice Comparison Test.

Participants

Twenty-nine native speakers of English aged 19 to 46 served as judges in this phase. While some of these judges were ESL teachers, the majority were working in unrelated areas.

Materials

The judges listened to the tape that resulted from the procedures of Phase 3.

The questionnaire eliciting their judgments contained the following four 7-point Likert scales: (1) Accent, (2) Ease of speech (speed, hesitations), (3) Ability to express ideas, (4) Overall ability in English (see Appendix D). The scale ranged from 1 (not at all native-like) to 7 (sounds like a native speaker). Our primary interest was in the judgment corresponding to item 3

(Ability to express ideas). It was thought it would be useful, however, to help the untrained judges eliminate considerations of accent and fluidity of speech from their judgment by explicitly focussing on accent and fluidity in separate scales.

Procedure

The judges were asked to listen to all 16 English samples of speech and rate them on the above-described four 7-point scales. The various judging sessions took place in different locations, but all sessions were conducted by the researcher herself following a uniform procedure. Except for one case where four ESL teachers judged the students at once, every judging session took place with only one judge at a time. Each judge was briefed on the importance of carefully listening to the voice before judging. Although they were shown a copy of the storyboard the voices on the tape were talking about, at no time were the judges given any specific directions as to what to look for in terms of speech characteristics. In order to ensure that judges understood how to use the scale, they were first asked to rate their own ability in French, their second language, with respect to telling the story depicted on the story board. After this the judges were encouraged to ask questions about the questionnaire, scales, experiment, etc. Fifteen judges heard the first version of the tape, the other fourteen heard the second version.

Phase 5 : Selection for final tape

As presented earlier, because the subjects taking the test were required to rate themselves against the voices on the tape, it was very important to carefully select such voices. A second factor which needed to be addressed was the novelty of the process of self-assessing against a set of voices. Seeing that learners may or may not have been accustomed to self-assessing that way, it was thought that including a "filler" or warm-up voice to begin with would offer them a chance to start thinking about the process before it actually counted. Finally, measures were taken to address the possibility of learners not following the instructions correctly (e.g. blindly rating themselves better or worse than all voice samples). These measures involved the use of 2 voices representing each level of ability (to provide a measure of consistency) as well as a second filler voice (native English speaker) to allow the subjects to experience at least one sample at the higher end of the scale. Responses to all eight voices were used in the FACETS analysis.

Procedure

The anglophone judges' ratings on item 3 of the questionnaire (Ability to express ideas) from Phase 4 were analyzed as follows.

First, the mean scores out of 7 and the standard deviations were computed for each voice sample. Six voice samples were identified from this analysis, the two with mean scores closest to positions 2, 4, and 6, with the

smallest standard deviations. These are designated as 2_1, 2_2, 4_1, 4_2, 6_1, and 6_2 in Table 3.2 along with their means and standard deviations. T-tests on the judges' scores on each pair of voices for each level (2_1 versus 2_2, 4_1 versus 4_2, and 6_1 versus 6_2) were conducted to determine if they were significantly different from one another despite being close to the same scale point (see Table 3.3, rows 1-3, for a summary). These t-tests indicated that the selected voices did not differ significantly from each other.

The next step was to determine that the voices rated as being 2's, 4's and 6's were significantly different from voices at other levels. For this, t-tests were conducted to compare ratings for the following comparisons: 2_1 versus 4_1, 2_1 versus 4_2, 2_2 versus 4_1, 2_2 versus 4_2, 6_1 versus 4_1, 6_1 versus 4_2, 6_2 versus 4_1, and 6_2 versus 4_2 (see Table 3.3, rows 4-11). These t-tests determined that the 2's differed significantly form the 4's and the 6's differed significantly from the 4's ($p<.001$ in all cases but one, and $p<.006$ in the remaining case).

Once these 6 voices were selected for the reasons mentioned above, two filler voices F_1 and F_2 were added. One filler voice (F_1) was a voice sample which had been rated as a 2 (not significantly different from 2_1 and 2_2 (see Table 3.3, rows 12 - 13). The second filler voice (F_2) was that of an anglophone (level 7). The final version of the tape contained these eight voices placed in the following quasi random order: 2-4-6-2-7-4-2-6.

Table 3. 2

Mean Rating and Standard Deviations for the Stimulus Voices.

Voice	Exact score	SD
2_1	2.321	1.156
2_2	2.103	1.012
F_1 Filler	3.071	1.386
4_1	3.586	1.119
4_2	3.552	1.055
6_1	6.000	1.089
6_2	5.786	1.197
F_2 Filler	6.586	0.825

Table 3. 3

T-tests for the Voices Selected for Voice
Comparison Test Tape.

Comparison	T (27)	p<
1. $6_1 - 6_2$ *	.938	n.s
2. $4_1 - 4_2$	.806	n.s
3. $2_1 - 2_2$	1.000	n.s
4. $2_1 - 4_1$	-4.629	.001
5. $2_1 - 4_2$	-3.338	.006
6. $2_2 - 4_1$	-7.389	.001
7. $2_2 - 4_2$	-6.817	.001
8. $6_1 - 4_1$	-10.773	.001
9. $6_1 - 4_2$	-9.451	.001
10. $6_2 - 4_1$	-5.333	.001
11. $6_2 - 4_2$	-6.058	.001
12. $F_1 - 2_1$	-1.000	n.s
13. $F_1 - 2_2$	-1.395	n.s

Phase 6 : Questionnaire design

The second language learners of English who were to use the Voice Comparison Test were given a special questionnaire to use. This questionnaire was divided into six sections.

Background information and consent

The first two sections of the questionnaire solicited background information about the learners and their consent in participating in the study (see Appendix E). These sections were written in the subjects' mother tongue (French) in order to ensure full understanding.

Storyboard

Next, there followed a copy of the storyboard which was used by the speakers recorded on tape. The learners were asked to examine the storyboard and to rate the level of difficulty of telling that story in French. This was done to ensure that they actually focussed their attention on the pictures in the storyboard.

Self-assessment scales

Next followed a French language description of the task they were to complete together with the 8 scales to be used for self-assessing, i.e. one scale

for each voice heard on the tape (see Appendix F). Each learner was asked to pay careful attention to the voices they heard, to imagine they had to tell the story on the storyboard in English, their second language, and to circle whether they would expect their own performance of telling the story to be worse than, as good as, or better than the performance of the person recorded on tape.

Motivation and attitude questionnaires

The last part of the questionnaire included questions designed to address subjects' motivation and attitude toward English, their second language (see Appendix G). The questions were French translations of those used in Blanche's (1988) study of self-assessment. Blanche had selected 7 questions from a larger questionnaire designed by Lyons (1982). The 7 selected questions were divided into two sections. The first 4 questions addressed the participants' motivation and learning efforts. Each question involved a three-point scale where 1 indicated a very lower identification with the motivational trait and 3 indicated a strong identification with the trait. The motivation score was obtained by summing across responses for all four questions, yielding a score that ranged from 4 to 12. The remaining 3 questions addressed the participants' attitude toward English as a second language and involved 3 five-point items, where 1 meant "not at all comfortable" with the situation described and 5 meant "very comfortable".

37

The attitude score was obtained by summing across responses for all three questions, yielding a score that ranged from 3 to 15.

Phase 7 : Test administration

The last phase involved administration of the self-assessment test. This involved presenting the stimulus tape with the 8 selected voice samples (from Phase 5) to a group of English language learners who used the questionnaire (Phase 6) to assess their own ability to express themselves in English .

<u>Participants</u>

The final group of participants involved in the testing process was a group of 149 students (57 males) at College de Sherbrooke. The student population is mostly French Québécois counting very few members of minority groups. All of them were between 16 and 43 years old (Mean age = 19).

<u>Materials</u>

The stimulus tape developed in Phase 5 and the questionnaire developed in Phase 6 were used in this phase of the research.

Procedure

The test was administered to classes of between 21 and 34 students at a time. The author administered the tests herself, with the teacher absent form the room. The purpose of the test was explained to the students as being a study to see whether students could accurately indicate their own level of ability in their second language when asked appropriate questions. All students were assured that the results of their test would remain confidential and would only be used for research purposes. Students were very cooperative and showed great interest in the project and its outcome.

One week after this first session, the same students were asked to rate on a 7-point Likert-type scale their (1) global ability in English, (2) ability to express ideas, (3) ability to read, (4) ability to understand, and (5) ability to write in English (see Appendix H). These data were gathered for comparison with the results obtained on the VCT. The data given in their ratings of their global ability were retained for analysis.

CHAPTER 4 : RESULTS

The principal results of concern were those derived from the ESL learners' self-assessment ratings based on their judgements about their own speaking abilities upon listening to the voice samples in the final listening tape (Phase 6, Chapter 3). These results concern four main issues: the reliability and validity of the self-assessment tool, the relationship between self-assessment ability to individual difference factors, and the relationship between self-assessment scores and more conventionally derived assessment scores. These results are presented below. The results from the earlier phases of the methodology have already been presented in Chapter 3 since they were directly concerned with the development of the stimulus materials for the VCT.

Note that data were dropped from analysis for participants whose mother tongue was not French, who did not complete all self-assessments appropriately (e.g. left some comparisons blank or circled both *worse than* and *better than* for a single voice sample), or whose score on the CEGEP's placement test was unavailable. As a result, the final number of participants whose data was analyzed fell to 135.

The ratings provided by the ESL learners were organized into a matrix in which each row represented data from a single subject and the columns represented the 8 voice samples heard on the final tape (voices designated

earlier as $2_1, 2_2, 4_1, 4_2, 6_1, 6_2$, F_1 and F_2. The data were in the form of -1 (worse than), 0 (as good as), +1 (better than). This data matrix was then submitted to the Rasch Model of analysis – version 2.75 of FACETS on a 486 DX66 PC (Linacre, 1994). This analysis yielded the following measures relevant to the four main issues of concern in this investigation.

First, each of the 8 voice samples was located by the analysis on a continuum ranging from approximately -5 to +5 logit units, where the lower end of the scale represented "easy" test items (that is, voice samples that many subjects indicated represented speaking abilities that they could equal or surpass) and where the higher end of the scale represented "difficult" test items (that is, voice samples that fewer subjects indicated represented speaking abilities that they could equal or surpass). These locations on the scale are shown in column 4 of Table 4.1.

Second, the analysis also placed each of the 135 ESL learners on the same scale of logit units, where a location at the lower end of the scale indicated a learner with a relatively lower level of self-assessed ability and a location at the higher end of the scale indicated a learner with a relatively higher level of self-assessed ability.

Third, the analysis also provided a model indicating the optimal fit of the various responses and locations for each stimulus voice on an ability continuum. The number of iterations required by the Rasch analysis to achieve the best fit in this experiment was 38. Finally, the analysis provided a

41

Table 4. 1

Test Voices Measurement Report.

Voice #	Rasch		Native
	Logit scale	Standard	Speaker
	Value	Error	Ratings
2_1	-3.86	0.28	2.321
2_2	-3.01	0.23	2.103
F_1	-2.14	0.20	3.071
4_2	-1.18	0.18	3.552
4_1	-1.02	0.18	3.586
F_2	2.96	0.19	6.586
6_1	3.99	0.23	6.000
6_2	4.27	0.24	5.786

measure of the degree to which each voice sample or subject fit into the model (infit coefficients).

Reliability of the self-assessment measures.

As discussed at the end of Chapter 2, reliability of the measures was considered in terms of the internal consistency of the data. This is given by two measures. First, internal consistency is reflected, in part, in the ease with which the data permitted the generation of a best fit model. This is given by the number of iterations required for the analysis to reach an optimal solution. In the present case this number was 38 iterations. Second, the Rasch analysis computed a measure of fit between the raw data and the model scores (the Rasch equivalent of Cronbach's Alpha). This measure of internal consistency for the placement of the subjects on the model scale was .81 (RMSE = .88). This number represents the reliability of a person's given score on the scale being the "right" one. The corresponding measure for the placement of the stimuli voices was .99 (RMSE = .22). Here, the given .99 reliability is a function of many students giving similar judgements when the voices were heard on the stimulus tape. Overall, these results indicate a high level of internal consistency.

Note here that despite the good fit of the stimulus voices onto the one model, not all responses to the voices fit onto the model easily. Of the 1112 responses (individual self-assessments), 16 responses were categorized as

being unexpected. This means to say that 16 of the responses given by a variety of different participants (13) to the different test items (different voice for each participant) were not what the Model would have expected.

Validity.

As explained in Chapter 2, validity of the Voice Comparison Test was addressed by comparing how native English speakers rated the voice samples and how the ESL learners implicitly rated the voice samples in the course of making judgements about themselves. For this analysis the mean rating scores from the 29 native English speakers for the 8 voices used on the final tape (2_1, 2_2, 4_1, 4_2, 6_1, 6_2, F_1 and F_2) were correlated against the locations on the logit scale for these same voices derived from the ESL learners' data as given by the Rasch analysis (see Table 4.1). The correlation obtained was very high and significant ($r=0.965$, $p< .001$).

Group and individual characteristics.

The next analysis investigated possible relationships between a measure of ability to assess one's own level of communicative ability and 5 different individual characteristics–learner ability level, sex, age, motivation, attitude toward English, and age at which the participant started to learn English.

Each participant's ability to consistently self-assess is indicated by their

44

infit coefficients. These infit coefficients indicate how difficult it was for the Rasch analysis to fit the person's self-assessment scores onto the model.

1. Level of Ability. The issue here was whether high ability learners were better able to consistently assess their own communicative ability than lower ability learners. In trying to determine how consistently ESL learners at different levels of achievement are able to self-assess, it was necessary to define a measure of "level of ability" for each participant. Three such definitions were found based on the data collected in this study.

The first measure—here called SCORE3—made use of the results participants obtained on the placement test used by the institution. Note that the placement test administered to students entering an ESL class at College de Sherbrooke was the TCAL-100. This test lasted between 45 and 60 minutes. The test included a questionnaire and an audio tape where directions were written and read in French. The time allotted to each question was dictated by the voice on the tape and students could not go back to any of the questions. There were 4 sections in the test, for a total of 100 questions (see Appendix I).

This subdivision of the placement test scores was a 3-category scheme in which category 1 represented the beginner level class, category 2 represented the intermediate-level class, and category 3 represented the advanced-level class. The mean infit coefficients for each of the three placement categories were 1.70, 1.00, and 0.71 respectively (see Table 4.2).

Table 4. 2

Mean Infit Coefficients for Participants by Class Level.

Level	N	Mean	SD
1	13	1.70	1.496
2	71	1.00	0.810
3	48	0.71	0.443

Given the uneven distribution of N across the various categories (ranging from 13 to 71), it was not possible to conduct an analysis of variance.

The second measure–here called SCORE5–was a refinement of SCORE3. Here the learners were divided into five categories instead of 3, based on the placement test scores. These 5 new categories were established by first taking the mean placement test score from each of the 3 class levels established by the institution (level 1 mean score = 26.1 (SD=7.4); level 2 mean score = 50.4 (SD=8.7); level 3 mean score = 82.1 (SD=6.9). To these means were added and subtracted one standard deviation, resulting in the following new category boundaries: level 1 = 0.0 to 33.8, level 2= 33.9 to 41.1, level 3= 41.2 to 60.3, level 4= 60.4 to 75.1, level 5= 75.2 to 95.0 The infit coefficients for these 5 categories are summarized in Table 4.3. As with the SCORE3 data, given the uneven distribution of N across the various categories (ranging from 1 to 56), it was not possible to conduct an analysis of variance.

The third way of determining the participants' ability level was to use the logit scores computed for each individual by the Rasch analysis. Thus each subject had an infit score and a logit score which placed him or her on an ability continuum. The correlation between these scores was r = -0.153, n = 135, n.s.

2. <u>Sex</u>. The mean infit score for females was .940, SD = .801 and for males it was .991, SD = .907. A t-test run on the infit scores for both sexes showed that men and women made equally consistent self-assessments:

47

Table 4. 3

Mean Infit Coefficients of Participants by Level (5)

Level	N	Mean	SD
1	1	0.700	0.000
2	4	2.125	1.457
3	18	1.350	1.297
4	54	0.993	0.755
5	56	0.731	0.541

48

t(133)= -0.327, n.s.

3. Age. The correlation between participants' age and their infit

coefficients was r = .142, n = 92, n.s.

4. Motivation. The data for this section as described in Chapter 3 were

based on the scores participants obtained on the motivation questionnaire. A

number of participants' data could not be included in this section because

they did not complete the motivation questionnaires adequately (final

N=127). The three reasons for eliminating some subjects' data were the

following: (1) participants did not complete all of the questions, (2)

participants misunderstood the way they were supposed to answer or (3) they

gave more than one answer to each question. The correlation between

participant motivation and infit coefficients was calculated at r = -.066, n.s.

5. Attitude. Data dealing with attitude were provided by the scores

students obtained on the attitude questionnaire.

As with the motivation results, the number of participants whose data

could be used in this analysis was reduced to N= 101 by incomplete

questionnaires. The final correlation obtained for the reduced size sample was

r= .004, n.s.

6. Age of first contact with English (learning context) Based on the

information gathered form the participants, the correlation between the age at

which participants started to learn English and their infit coefficients was

calculated at r = .481, p<.001.

4. Relation between self-assessments based on the Rasch analysis and conventional placement measures.

Data here include two comparisons. The first set of comparisons was between the scores obtained on the Rasch scale and those obtained on the various classifications of the institution's test scores. The correlation between the Rasch scores and the original placement test scores (max = 100) was: r=.588, p<.001. The correlation between the Rasch scores and the SCORE3 classification of the placement test scores was: r= .541, p<.001. The correlation between the Rasch scores and the SCORE5 classification of the placement test scores was: r= .475, p<.001. The second comparison was between the Rasch scores and the means obtained on the Likert-type scales. This correlation was r=.542, p<.001.

CHAPTER 5 : DISCUSSION

This thesis aimed to see whether a test could be developed that would reliably and validly determine ESL learners' assessment of their own ability to communicate in their second language. The test under consideration was a novel self-rating procedure in which second language users compared their own ability to express ideas in the second language with the ability they attributed to other speakers on the basis of listening to voice samples from those speakers. A test was developed to do this and was evaluated with respect to the following four research questions. First, did the test yield reliable self-assessment ratings? Second, did the test yield valid ratings? Third, was the ability to self-assess consistently, as measured in this test, systematically related to certain subject characteristics? Finally, how did the results produced by the Voice Comparison Test compare with results from a more conventional self-assessment test? These issues are now discussed in turn.

Reliability

In a study conducted on the reliability of a specific self-assessment tool, Davidson and Henning (1985) urged that "every precaution be taken and every tool be employed to ensure that rating scales be applied in the most accurate, meaningful and readily interpretable manner possible" (p. 164). As part of such precautions, they attempted to apply a specific Item Response

Theory (IRT) model called the Rasch model to self-assessments scores generated by the a specific self-assessment tool. Their conclusion was that the results generated by the tool did not answer all four of the requirements of the Rasch model. First, as indicated by the high misfit scores, the authors believed the data to be acting along two dimensions rather than the one required by the model. Second, the more–or–less comparisons for persons self-assessing were difficult at the lower end of the scale. Third, support for the concept of linear magnitude was not met to the extent desirable as some of the points along the continuum were in fact too close together to be distinguishable. The requirement of determining a replicable unit of measurement is the only one which was found to be appropriate and in accordance with the conditions and assumptions of the Rasch Model. As described in Chapter 2, the data presented in this study met these minimum requirements suggesting that the Rasch Model could be used to analyze the data gathered with the VCT.

One indicator of consistency in the data is the number of iterations required by the analysis to discover a best fitting model. In the present case this was 38 iterations. According to Upshur (Personal communication, February, 1997), this can be considered to indicate that the data were not problematic or inconsistent as the default number of iterations automatically allowed by the software to find a solution to fit the data is 100. Insofar as this indicates that the data are free of noise and anomaly that would otherwise

make it difficult to find a solution to the demands of the modelling, we can conclude that they reflect a high degree of internal consistency. A stronger indicator that the data were free of noise, and in that sense internally consistent and reliable, was the high degree of fit that was found between model and original data.

The ways in which data might depart from perfect consistency can be illustrated by the following examples. It is possible, for example, that some subjects might provide inconsistent responses of the sort where they report their ability to express themselves in English to be worse than a speaker represented by voice sample 2_1 (a weak speaker) but better than the speaker represented by voice sample 4_1 (a strong speaker). While no such blatant contradictions occurred in the data there were cases where subjects rated themselves as similar to one speaker at a given level (e.g., 4_1) but not the other (e.g., 4_2). The Rasch model accommodates such inconsistencies by allowing for the possibility that the two voice samples in question are not really equivalent (e.g., 4_1 and 4_2 lie at different points on the ability continuum). By making these adjustments the Rasch model provides us with a more accurate picture of the relative positions of the stimuli and is able to discover a representation of the data that reflects more internal consistency than would be the case if the voice samples were initially fixed at points along the continuum without taking into account how subjects perceived them.

As described earlier, after 38 iterations the analysis was able to create a

model that placed voice samples and subjects on an ability continuum onto which only 16 out of 1112 responses did not fit easily. This suggests there was a very high level of consistency (less than 2% of the data being judged inconsistent). Indeed, it is worth noting that an internal consistency reliability of .81 for an eight-item test is unusually high. A rule-of-thumb for a 10-item quiz is approximately .50.

Table 4.3 shows that the voices used on the tape in fact could be entered onto the scale relatively easily because their reliability scores were high.

The Rasch analysis of the learner's placements on the ability scale suggests no natural breaks in the sample group. The distribution of the participants does not naturally fall into three separate categories, which means that the separation of participants into three separate classes within their school is artificial. Of course, for practical reasons, divisions need to be made for classes to be created. For this to be done properly, however, careful analysis of the abilities associated with each score would need to be made for any division into sub-groups to be meaningful. Neither the test which was used by the school nor this self-assessment tool provides a basis for this division.

Validity

The second question asked in this study was whether or not there was evidence for the validity of the ability scale produced by this self-assessment

54

tool. As suggested in the results section, strong evidence for construct validity can be potentially obtained by considering the actual procedure used to obtain the anchor points (the stimulus voices representing different ability levels) on the scale itself.

There was a high and significant correlation ($r=.965$, $p<.001$) between the ratings of ability to express ideas given to the stimulus voices by native English speaking judges and those that emerged from the Rasch analysis of the responses to the same voices by the second language learners of English. This indicates that the two groups were using a similar reference system in perceiving and making judgments about the voices. To the extent that this indicates that the ESL learners perceived and used the voice samples for making judgments in a way that was highly consistent with how native English speakers perceived the voice samples, we can be confident that the ability scale produced by the Rasch analysis is valid and that the learners are validly located on this scale.

Note that there is another sense of validity which future research on the Voice Comparison Test could address. This is a comparison of how well the ESL learners' self-assessments of communicative ability matched their "true" communicative ability as indexed by some objective measure. Such a comparison would be an important step to take but it was not possible to do this within the framework of the present thesis. Nevertheless, it was felt that a match between the ESL learners' indirect perceptions of the voice samples

that emerged from an analysis of their self-assessments and the direct perceptions of native English speakers would speak to the issue of validity in a fundamental way.

Learner characteristics and consistency in self-assessment

Having determined that the scale produced by the Rasch analysis from the self-assessment ratings provided by the ESL learners meets certain basic criteria for reliability and validity, we can now consider some interesting questions about what factors might influence performance on the VCT. In this investigation we considered the possibilities that level of ability, sex, age, motivation, attitude, and age of first contact with English might be relevant factors. The method of analysis was to test the significance of an association between infit coefficients and measures of the participant characteristic under consideration. It will be recalled that the infit score provides a measure of the degree to which a given subject's data fit easily into the model produced by the Rasch analysis. It is assumed here that if a participant had difficulty in using the self-assessment scale then that participant's ratings would not be as consistent as they might otherwise be. (In an extreme case, for example, the participant might produce random responses with numerous inconsistencies emerging with respect to the eight stimulus voice samples.)

The first of these characteristics considered was ability level. It was thought that perhaps more advanced learners would be in a better position to

judge their own ability than would be less advanced learners. In looking at the level of ability of the learners as determined by either the score out of 100 they obtained on the school's placement test, the class division of their scores (beginner, intermediate, advanced), or the scores they obtained on the Rasch scale, there did not appear to be any difference at all in the consistency with which people self-assessed.

A number of authors have suggested that there are clear distinctions which exist between beginner, intermediate, and advanced level students when looking at their ability to self-assess. The two main contentions in the literature about level of ability with regards to ability to self-assess are that intermediate learners behave differently in self-assessment circumstances (Berkoff 1985, Blanche, 1988, 1989, von Elek, 1982), and the beginners overestimate their abilities while advanced learners underestimate (Blanche, 1990, Janssen- van Dieten, 1989, Heilenman, 1990, Leblanc & Painchaud, 1985, and von Elek, 1982). The present results neither confirm nor disconfirm these conclusions. However, the results do show that if participants were in fact underestimating, overestimating, or self-assessing correctly, they were doing so consistently.

Sex

The t-tests of infit scores differentiated by subjects' gender indicated that men and women did not significantly differ in how consistent they were

in the way they self-assessed against the various voices on the stimuli tape. This result should not be interpreted to mean that neither men nor women overestimate or underestimate their ability in the second language. What these numbers indicate is that if they are in fact overestimating or underestimating, their abilities, they are doing so in a very consistent fashion. In order to determine whether or not they are in fact overestimating or underestimating, it would be necessary to have some "objective" measure of the participants' ability in the second language (using some other criterion beside self-assessment) so as to then make comparisons with the score attributed to them by the Rasch Model.

Age, motivation, and attitude

In looking at the correlation between infit coefficients of participants and means of age, motivation or attitude, it is clear that no significant relationship can be said to exist between them. In the case of age, one could claim that a lack of correlation is an indication that between younger and older adults there does not seem to be a difference in reliability and consistency in self-assessment.

In the analyses of motivation and attitude, the lack of significant correlations with the infit coefficients could be interpreted as an indication that student motivation and attitude do not have an impact on one's ability to self-assess. Researchers such as Anderson (1982), Bachman (1991), Blanche

(1988, 1990), and Wesche, et al. (1990) suggested that learner motivation and attitude significantly increased or decreased (as the case may be) the ability of learners to self-assess.The present data do not support that.

Note, however, that in looking more closely at the data in this study, it could be suggested that there were problems with the measures of motivation and attitude. Because the questionnaire was an adaptation of questionnaires used originally with English speakers and not French speakers, it is possible that the validity of the questionnaires was diminished in the process of translation. No particular verification of the value of the questionnaire was done prior to the study. This means that even a strong positive correlation between the scores on these questionnaire and the test scores, had this been found, would have had to be further investigated.

Age at which participant started to learn English

There was a significant positive correlation ($r = 0.481$, $p<.001.$) between the age at which participants started to learn English and their ability to self-assess as indexed by the infit score.

In an attempt to determine whether it was the age at which students started to learns English which correlated well with their ability to self-assess consistently or the number of years of practice in using the language, we tried to factor out the effect of the number of years. This factoring out left very little data with which to work. A closer look at the nature of the data revealed that

except for 5 more mature students, the participants all started learning

English at the ages of 9 or 10. Furthermore, the mature students in question

did not, in fact, self-assess as consistently (higher infit scores) as did the rest of

the participants. Such characteristics of the data did not make factoring out

feasible.

Other ways of classifying individuals

We were also interested to see how the voice comparison test

developed in the present study compared to institutional placement tests in

terms of classifying students according to ability.

As was reported earlier, the correlations between the Rasch scores

derived from the voice comparison test and the scores obtained on the

school's placement test varied between r=0.475 (p<.001) and r=.588 (p<.001)

depending on the way the institutional scores were divided (original score

versus SCORE3 versus SCORE 5). Such correlations cannot, in principle,

validate the VCT since there is no validation evidence for the institutional

placement test itself. Nevertheless, such correlations can contribute to the

discussion about the validity of the voice comparison test.

The first thing these moderate yet highly significant correlations

between the two sets of scores indicate is that, to a certain degree, both of these

tests measure the same thing. These correlations indicate that there is perhaps

as much as 35% shared variance between the Voice Comparison Tool ratings

and the institutional placement test results.

The correlation found between the Rasch scores and the scores obtained with the Likert-type 7-point scale was 0.542 ($p<0.001$). This indicates that there is only 25 % shared variance in the two self-assessment tests. Of course, the unshared variance is presumably due to differences between the tasks. The VCT is highly focussed and, as we have seen, produces results that correspond to native speakers' judgements of the stimulus voices. The Likert-scale self-rating task may elicit responses that are somewhat more variable due to factors discussed in the introduction (reliance on memory, interpretation of what it means to be able to express oneself, selection of different reference points by different subjects, etc.)

Conclusions

The results presented in this thesis have indicated that the Voice Comparison Test is a promising method for assessing ability in a second language. In addition to meeting the criteria of ease of administration and practicality discussed in the introduction to this thesis, the high degree of consistency of the assessments provided by the measure and the evidence for validity justify such promise.

The investigation reported in this thesis has also shown that the Rasch Model of measurement as a means of analysis for the self-assessment data provides variables (e.g. infit scores) which could, in future, be used in

research contexts to provide interesting information about learners and their abilities in their second language. Based on the findings obtained in this study, various recommendations can be made for future research.

First, in future replications of this study, one might choose to include more voices on the stimulus tape so as to determine whether or not the reliability of the scale established would, in fact, increase. In doing this it might be interesting to find out what the optimal number of voices needed on the tape is before the reliability of the tool stabilizes or even decreases.

A second interesting avenue for future research might be to determine the link which exists between each level on the scale and actual ability in the second language as determined by independent measures of student ability. The earlier suggestion to record the subjects performing the story telling task themselves and having these judged by native speakers might be a way of providing such a measure of subject ability.

Third, it might be interesting to investigate whether the infit scores generated by the Rasch Model act as good predictors of student success in class. Through investigation it might be possible, for instance, to determine if subjects having lower infit scores are more sensitive to important language use variables and whether they make more progress in their course as a result of this increased sensitivity.

Finally, it would be interesting to investigate the possibilities of increasing the generalizability of findings by studying the use of the tool in

different settings. The tool could be tried: (1) with subjects from various language groups, (2) with subjects from various age groups, or (3) in different testing contexts (e.g. using the tool as a means of measuring progress as opposed to starting ability). A similar type of tool could also be designed to measure ability in other skill domains. For example, in measuring reading skill development it might be useful for researchers to determine the level of their participants by having them rate their reading skill in relation to sample passages of different levels of difficulty.

References

Aitken, K. G. (1977). Using Cloze procedure as an overall language proficiency test. TESOL Quarterly, 11 (1), 59-67.

Alderson J. C. (1986). Innovations in language testing. In Portal, M. (Ed.), Innovations in language testing: Proceedings of the IUS/NFER conference, April, 1985. (pp. 93-108). Windsor: NFER-Nelson.

Anastasi, A. (1989). Ability testing in the 1980s and beyond: Some major trends. Public Personnel Management, 18 (4), 471-485.

Anderson, P. (1982). Self-esteem in the foreign language: A preliminary investigation. Foreign Language Annals, 15, 109-115.

Bachman, L. F. (1985). Performance on Cloze tests with fixed-ratio and rational deletions. TESOL Quarterly, 19 (3), 535-513.

Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford, England: Oxford University Press.

Bachman, L. F. (1991). What does language testing have to offer? TESOL Quarterly, 25 (4), 671-704.

Bachman, L. F., & Palmer, A. (1988). The construct validation of self-ratings of communicative language ability. Language Testing, 6, 14-29.

Bachman, L. F., & Savignon, S. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. Modern Language Journal, 70 (4), 380-390.

Berkoff, N.A. (1982). Testing oral proficiency: A new approach. In Lee, Y., Fok, A., Lord, R., & Low, G. (Eds.), New directions in language testing (pp. 93-99) Oxford, England: Oxford University Press.

Blanche, P. (1988). Self-assessment of foreign language skills: Implications for teachers and researchers. ELC Journal, 19 (1), 75-93.

Blanche, P. (1990). Using standardized achievement and oral proficiency tests for self-assessment purposes: The DLIFLC study. Language Testing, 7, 202-229.

Blanche, P. & Merino, B. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. Language learning, 39 (2), 313-337.

Burt, M., & Dulay, H. (1978). Some guidelines for the assessment of oral language proficiency and dominance. TESOL Quarterly,12 (2), 177-192.

Clark, J. L. D., & Clifford, R.T. (1988). The FSI/IRL/ACTFL proficiency scales and testing techniques. Studies in Second Language Acquisition, 10, 129-147.

Clark, J. L. D., & Lett, J. (1988). A Research Agenda. In Lowe, P., and Stansfield, C. W. (Eds.), Second language proficiency assessment: Current issues. Language in education: Theory and practice (pp. 53-82). New York: Prentice-Hall.

Davidson, F., & Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. Language Testing, 2 (2), 164-179.

Dickinson, L. (1987). Self-instruction in language learning. Cambridge, England: Cambridge University Press.

Douglas, D. (1995). Developments in language testing. Annual Review of Applied Linguistics, 15, 167-187.

Farhady, H. (1979). The disjunctive fallacy between discrete-point and integrative tests. TESOL Quarterly, 13 (3), 347-357.

Ferguson, N. (1978). Notes and discussion: Self-assessment of listening comprehension. International Review of Applied Linguistics in Language Teaching, 16 (2), 149-156.

Hale, G., Stansfield, C., Rock, D., Hicks, M., Butler, F., & Oller, J. (1989). The relation of multiple-choice Cloze items to the Test of English as a Foreign Language. Language Testing, 6 (1), 47-76.

Hamayan, E. V. (1995). Approaches to alternative assessment. Annual Review of Applied Linguistics, 15, 212-226.

Heilenman, K. (1990). Self-assessment of second language ability: The role of response effects. Language Testing, 7, 174-201.

Holec, H. (1981). Plaidoyer pour l'auto-évaluation. Français dans le Monde, 21, 15-23.

Janssen-van Dieten, A.-M. (1989). The development of a test of Dutch as a second language: The validity of self-assessment by inexperienced subjects. Language Testing, 6 (1), 30-46.

Laesch, K. B., & Van Kleek, A. (1987). The Cloze test as an alternative measure of language proficiency of children considered for exit from bilingual education programs. Language Learning, 37 (2), 171-190.

Leblanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. TESOL Quarterly, 19 (4), 673-687.

Lewkewicz, J. A., & Moon, J. (1985). Evaluation: A way of involving the learner. In Alderson, J. C. (Ed.), Lancaster practical papers in English language education. Vol. 6. Evaluation (pp. 45-80). Oxford, England: Pergamon Press.

Linacre, J. M. (1989-1993). Users guide to FACETS: Rasch Measurement Computer Program. Chicago, IL: Mesa Press.

Linacre, J. M. (1994). FACETS (version 2.75) [computer software]. Chicago, IL: MESA Press.

Macnamara, J. (1967). The bilingual's linguistic performance – a psychological overview. Journal of Social Issues, 22 (2), 58-77.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62 (3), 229-258.

Moss, P. A. (1994). Can there be validity without reliability? Educational Research, 23 (2), 5-12.

Mullen, K. A. (1978). Direct evaluation of second-language proficiency: The effect of rater and scale in oral interviews. Language Learning, 28 301- 308.

Oscarson, M. (1980). Approaches to self-assessment in foreign language learning. Strasbourg. Council of Europe. Oxford, England: Pergamon Press.

Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. Language Testing, 6 (1), 1-13.

Painchaud, G., & Leblanc, R. (1984). L'auto-évaluation en contexte scolaire. Etudes de Linguistique Appliquée, 56, 88-98.

Promoudou, L. (1995). The backwash effect: From testing to teaching. ELT Journal, 49 (1), 3-25.

Raasch, A. (1978). "S'auto-évaluer" un néologisme ? L'auto-évaluation dans un cours de français langue étrangère. Le Français dans le Monde, 49, 63-67.

Rossi, G. J. (1983). Foreign language proficiency testing: A selected, annotated guide to standardized instruments, rating scales, literature. The French Review, 56 (6), 840-858.

Ryan, E. B. & Caranza, M. A. (1975). Evaluative reactions of adolescents towards speakers of standard English and Mexican and American accented English. Journal of Personality and Social Psychology, 31, 855-863.

Segalowitz, N. (1976). Communicative incompetence and the non-fluent bilingual. Canadian Journal of Behavioral Science, 8 (2), 122-131.

Upshur, J. A. (1975). Objective evaluation of oral proficiency in the ESL classroom. In Palmer, L. and Spolsky, B. (Eds.), Papers on language testing: 1967-1974. Washington D.C.: TESOL.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. ELT Journal, 49 (1), 3-12.

von Elek, T. (1982). A test of Swedish as a second language: An experiment in self-assessment. In Lee, Y., Fok, A., Lord, R., and Low, G. New directions in language testing (pp. 47-57). Oxford, England: Oxford University Press.

Wall, D., Clapham, C., & Alderson, C. (1994). Evaluating a placement test. Language Testing, 11 (3), 321-344.

Wangsotorn, A. (1980). Self-assessment in English skills by undergraduate and graduate students in Thai universities. In J. Read (Ed.), Directions in language testing (pp. 240-260). Singapore: Singapore University Press.

Wesche M.B., Morrison, F., Ready, D., & Pawley, C. (1990). French immersion: Post-secondary consequences for individuals and universities. Canadian Modern Language Review, 46, 430-451.

Whiteson, V. (1981). Foreign Language Testing: A current View. ELT
Journal, 35 (3), 345-352.

Wright, B., & Stone, M. (1979). Best test design: Rasch measurement.
Chicago, IL: Mesa Press.

APPENDIX A

Consent Form

This is to state that I agree to participate in a research project being conducted by Edith Dandenault under the supervision of Prof. Elizabeth Gatbonton of the TESL Centre at Concordia University.

I have been informed that the broad purpose of the research is to investigate the nature of second language speaking abilities.

I understand that I will be asked to do the following: I will be asked to look at a cartoon picture and to tell, in French and in English, the story that it shows. I understand that my telling of the story will be recorded. I also understand that this recording will later be played to a group of listeners in follow-up research.

I understand that the data generated by this study and in follow-ups which may use my tape recording will be submitted for publication. I also understand that the tape recording or portions of it may be incorporated into a testing instrument aimed at identifying levels of skill in a second language. There is the possibility that this testing instrument may be distributed widely for research purposes and possibly distributed commercially to institutions involved in language testing, training or research.

In signing this consent form I hereby waive all rights concerning the above uses that may be made of my recording, with the understanding that my identity will be protected at all times.

I understand that my participation is expected to last about half an hour or less and that I will be paid $5.00 for my participation.

I understand that I am free to withdraw my consent and discontinue my participation at any time without negative consequences.

I understand that my participation in this study is confidential, that is, my identity will not be disclosed in any written or oral reports of this research.

I understand that the data of this study will be presented in a research paper.

I understand that I may keep a copy of this consent form and that if I wish, I may receive a copy of the written report of this study (please allow several months). This can be obtained from Dr. Elizabeth Gatbonton, TESL Centre, Concordia University, 1455 de Maisonneuve Blvd., West, Montreal Quebec H3G 1M8.

I HAVE CAREFULLY STUDIES THE ABOVE AND UNDERSTAND THIS AGREEMENT. I FREELY CONSENT AND AGREE TO PARTICIPATE IN THIS STUDY.
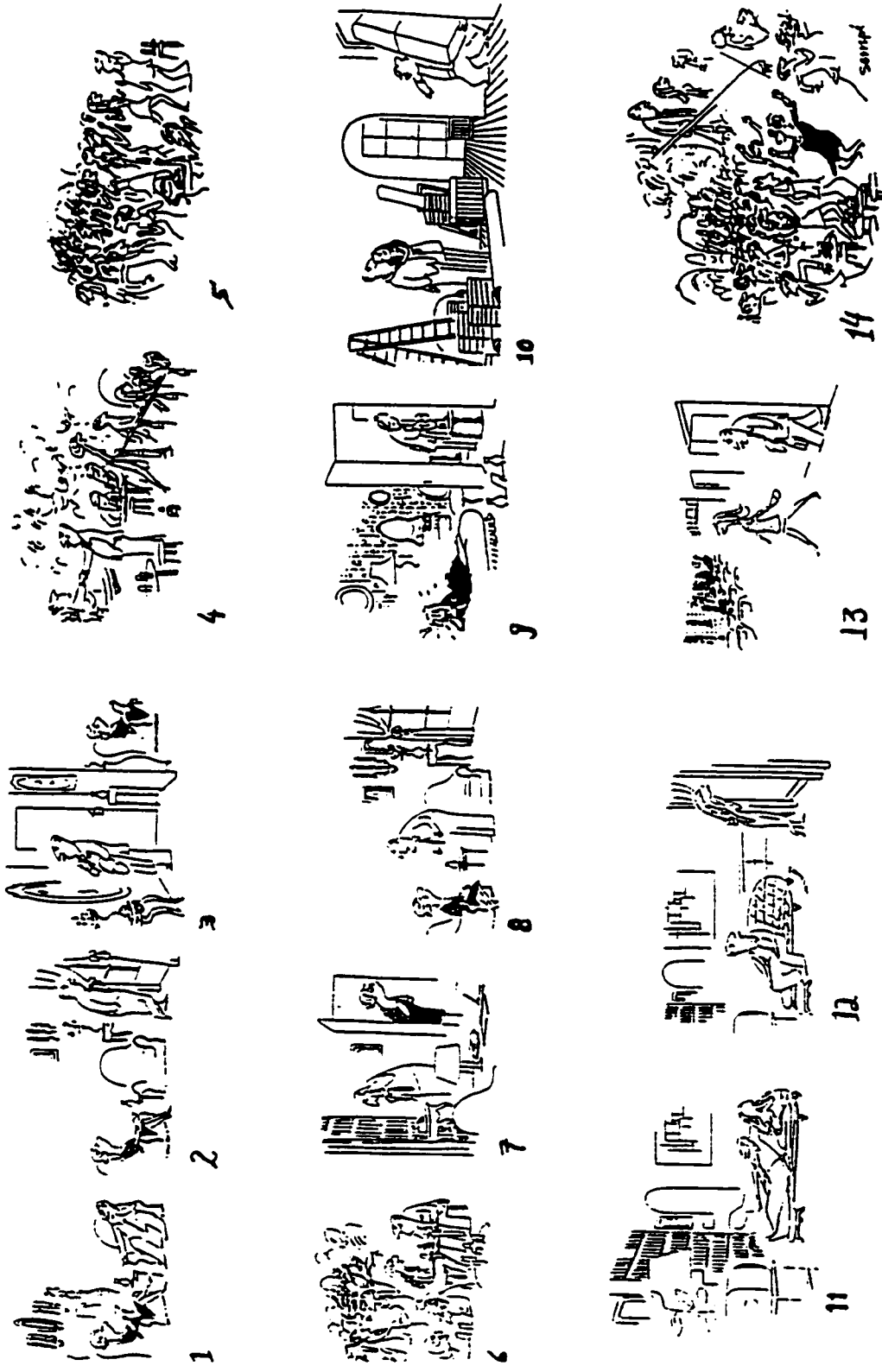
NAME (print) _____

SIGNATURE _____

DATE: _____

APPENDIX C

Évaluer chaque voix selon les caractéristiques ci-dessous sur l'échelle de 1 à 7. Encercler votre choix.

_____

Instruite	7	6	5	4	3	2	1	Pas instruite
Riche	7	6	5	4	3	2	1	Pauvre
Accomplie	7	6	5	4	3	2	1	Pas accomplie
Intélligente	7	6	5	4	3	2	1	Pas intélligente
Fiable	7	6	5	4	3	2	1	Pas fiable
Gentille	7	6	5	4	3	2	1	Pas gentille
Sympathique	7	6	5	4	3	2	1	Pas sympathique
Généreuse	7	6	5	4	3	2	1	Pas généreuse
Honnête	7	6	5	4	3	2	1	Pas honnête
Travaillante	7	6	5	4	3	2	1	Pas travaillante

APPENDIX D

Speaker #1 Rate the speaker in terms of the following:

1. How native-like is his accent when speaking English?

not at all native-like	1	2	3	4	5	6	7	sounds like a native speaker

2. How at ease at speaking English is this speaker (as reflected in speaking speed, hesitation, etc.)?

Extremely ill at ease	1	2	3	4	5	6	7	as comfortable as a native speaker

3. What is his ability to express ideas in English?

cannot express ideas clearly at all	1	2	3	4	5	6	7	can express ideas as well as a native speaker

4. All things considered, how would you rate this speaker's ability in English?

no ability at all	1	2	3	4	5	6	7	native-like ability

APPENDIX E

Par la présente je vous invite à participer à une activité qui cherche à
déterminer si les gens peuvent reconnaître leur propre niveau d'habileté
dans leur deuxième langue. L'activité au complet ne prendra que quelques
minutes.

Il est à noter qu'aucun renseignement obtenu durant cette activité ne sera
rendu publique de manière à divulguer votre identité. De plus, la
confidentialité de tous ces renseignements sera en tout temps respectée.
Enfin, les enseignants de cette institution n'auront jamais accès à
l'information contenue dans ce feuillet.

Je consens à participer: _____   _____

                                    Nom en lettres moulées   Signature

_____

_____

Nom: _____   Matricule: _____

Âge: _____   Sexe: F   M

Langue maternelle: _____   Langue seconde: _____

Autre(s) langue(s): _____

Âge auquel tu as commencé à apprendre l'anglais: _____

Pour les items ci-dessous, coche la réponse la plus appropriée pour toi.

Dans une semaine normale, est-ce que tu:

lis en anglais: _____ _____ _____ _____ _____
écris en anglais: _____ _____ _____ _____ _____
regardes la télé
    en anglais: _____ _____ _____ _____ _____
écoutes la radio: _____ _____ _____ _____ _____
parles en anglais: _____ _____ _____ _____ _____

73

1. Imagine qu'il faut que tu me racontes l'histoire en anglais.

2. Prépare-toi à entendre huit (8) personnes différentes raconter cette histoire. Tu n'entendras pas toute l'histoire de chaque personne, mais plutôt un extrait. Ensuite, après chaque extrait, dis-moi si tu penses que tu raconterais cette histoire:

   - **MOINS BIEN** que la personne que tu viens d'entendre.
   - **AUSSI BIEN** que la personne que tu viens d'entendre.
   - **MIEUX** que la personne que tu viens d'entendre.

3. Souviens-toi qu'il n'y a pas de "bonne" ou "mauvaise" réponse, l'important c'est d'être *le plus précis* possible.

   Après l'écoute de chaque extrait, ENCERCLE la réponse qui décrit le mieux comment tu raconterais la même histoire en <u>anglais</u> en comparaison avec la voix entendue.

   Je raconterais l'histoire: MOINS BIEN   AUSSI BIEN   MIEUX   que VOIX #1

   Je raconterais l'histoire: MOINS BIEN   AUSSI BIEN   MIEUX   que VOIX #2

   Je raconterais l'histoire: MOINS BIEN   AUSSI BIEN   MIEUX   que VOIX #3

   Je raconterais l'histoire: MOINS BIEN   AUSSI BIEN   MIEUX   que VOIX #4

   Je raconterais l'histoire: MOINS BIEN   AUSSI BIEN   MIEUX   que VOIX #5

   Je raconterais l'histoire: MOINS BIEN   AUSSI BIEN   MIEUX   que VOIX #6

   Je raconterais l'histoire: MOINS BIEN   AUSSI BIEN   MIEUX   que VOIX #7

   Je raconterais l'histoire: MOINS BIEN   AUSSI BIEN   MIEUX   que VOIX #8

Pour chacune des questions ci-dessous, coche la réponse la plus appropriée pour toi.

1. Comparativement à la moyenne de mes amis(es), je pense que:

_____ j'étudie l'anglais PLUS que la plupart d'entre-eux

_____ j'étudie l'anglais AUTANT que la plupart d'entre-eux

_____ j'étudie l'anglais MOINS que la plupart d'entre-eux

2. Je pense aux mots et aux idées que j'apprends en anglais

_____ presque jamais

_____ de temps à autre

_____ fréquemment

3. Considérant les efforts que je mets à apprendre l'anglais, je peux honnêtement dire que

_____ je vais l'apprendre par chance ou intelligence puisque je ne travaille presque pas

_____ j'en fais juste assez pour arriver

_____ j'essaie vraiment d'apprendre la langue

4. Quand j'aurai terminé d'étudier ici je vais probablement

_____ ne faire aucun effort pour me rappeler de la langue

_____ continuer d'améliorer mon anglais

_____ essayer d'utiliser la langue le plus souvent possible

Pour les questions de 5 à 7, fais un x vis-à-vis le tirait le plus juste.

5. Lorsque je parle anglais à un étranger qui le parle bien, je me sens:

très confortable	__ __ __ __ __	pas du tout confortable
très confiant(e)	__ __ __ __ __	pas du tout confiant(e)
très gêné(e)	__ __ __ __ __	pas du tout gêné(e)

6. Lorsque je parle à mes enseignants en anglais, je me sens:

très confortable	__ __ __ __ __	pas du tout confortable
très confiant(e)	__ __ __ __ __	pas du tout confiant(e)
très gêné(e)	__ __ __ __ __	pas du tout gêné(e)

7. Lorsque je parle en anglais à des amis(es) qui le parle bien, je me sens:

très confortable	__ __ __ __ __	pas du tout confortable
très confiant(e)	__ __ __ __ __	pas du tout confiant(e)
très gêné(e)	__ __ __ __ __	pas du tout gêné(e)

Nom: _____

D'après toi, où te situes-tu sur les échelles suivantes :
(encercle le numéro qui correspond à ton niveau)

1. Habilité globale en anglais :

1	2	3	4	5	6	7
Aucune						Comme
habileté						une personne
						anglaise

2. Habileté à exprimer tes idées oralement en anglais :

1	2	3	4	5	6	7
Aucune						Comme
habileté						une personne
						anglaise

3. Habileté à comprendre ce que tu lis en anglais :

1	2	3	4	5	6	7
Aucune						Comme
habileté						une personne
						anglaise

4. Habileté à comprendre ce que tu entends en anglais :

1	2	3	4	5	6	7
Aucune						Comme
habileté						une personne
						anglaise

5. Habileté à écrire ce que tu veux en anglais :

1	2	3	4	5	6	7
Aucune						Comme
habileté						une personne
						anglaise

DO NOT WRITE

IN THIS BOOKLET !

NE RIEN INSCRIRE

DANS CE CAHIER !

COLLEGE DE SHERBROOKE
LANGUES MODERNES

TEST DE PRÉ-CLASSEMENT D'ANGLAIS
LANGUE SECONDE AU NIVEAU COLLÉGIAL
SECTION COMPRÉHENSION.

Le test qui vous sera administré dans les quelques quarante minutes qui suivent, est un test de pré-classement d'anglais, langue seconde, qui est destiné aux étudiants québécois francophones de niveau collégial. Les résultats de ce test seront utilisés par vos professeurs d'anglais afin de les aider à déterminer le niveau de cours qui correspond à votre niveau de connaissance de l'anglais.

Le test comporte quatre parties totalisant cent questions; la durée du test est approximativement de 35 à 40 minutes. Chacune des parties est précédée de directives que vous pouvez lire sur ce questionnaire pendant qu'elles vous seront lues. Votre professeur n'aura pas à intervenir puisque l'enregistrement que vous entendez comporte à chacune des questions. nous vous suggérons de suivre le rythme de travail que vous impose cet enregistrement; il vous laisse suffisamment de temps pour répondre à toutes les questions sans pour autant vous permettre de revenir en arrière sur des questions auxquelles vous n'auriez pas eu le temps de répondre.

Nous vous prions de consacrer le plus grand soin à votre apprentissage de l'anglais, langue seconde, en nous aidant à connaître votre niveau réel de compétence pour vous orienter ensuite vers le cours d'anglais qui correspond le mieux à vos besoins.

Phonologie

Directives: Pour les 20 premières questions, vous entendrez une phrase qui vous
                     sera répétée deux fois; choisissez sur votre questionnaire la phrase qui
                     correspond le plus exactement à ce que vous aurez entendu. Si la
                     phrase que vous avez entendu ne correspond ni à l'une ni à l'autre des
                     deux premières phrases que vous pouvez lire sur votre questionnaire,
                     choisissez la réponse " Neither of the above"; au contraire, si la phrase
                     que vous avez entendu correspond à la fois aux deux premières
                     phrases que vous pouvez lire sur votre questionnaire, choisissez alors
                     la réponse "Both of the above".

                     Now listen carefully to item no. 1.

1.   a) Can you test it?
       b) Can you paste it?
       c) Neither of the above.

2.   a) Clean air smells good.
       b) Clean hair smells good.
       c) Neither of the above.

3.   a) He took a bat.
       b) He took a bath.
       c) Both of the above.

4.   a) He was panting.
       b) He was painting.
       c) Neither of the above.

5. a) George caught the rope.

   b) George cut the rope.

   c) Neither of the above.

6. a) Don't sleep on the deck.

   b) don't slip on the deck.

   c) Neither of the above.

7. a) It looks like a happy phase.

   b) It looks like a happy face.

   c) Neither of the above.

8. a) She threads needles.

   b) She dreads needles.

   c) Both of the above.

9. a) Hold the cart.

   b) Hold the card.

   c) Neither of the above.

10. a) The pad was crooked.

    b) The path was crooked.

    c) Neither of the above.

11. a) Look at the water fall.

    b) Look at the waterfall.

    c) Both of the above.

12. a) You have to mail boxes at the post office.

    b) You have two mail boxes at the post office.

    c) Neither of the above.

13. a) Bill is coming home.

    b) Bill is coming home!

    c) Bill is coming home?

14. a) We like Dick.
    b) We liked Dick.
    c) Both of the above.

15. a) They watch television.
    b) They watched television.
    c) Neither of the above.

16. a) He has goldfish in the tank.
    b) He has gold fish in the tank.
    c) Neither of the above.

17. a) I want two race horses.
    b) I want to race horses.
    c) Neither of the above.

18. a) He's scared of mice?
    b) He's scared of mice.
    c) He's scared of mice!

19. a) He had a nice boat.
    b) He had an iceboat.
    c) Neither of the above.

20. a) You asked Tim.
    b) You ask Tim.
    c) Neither of the above.

Première section

Directives: Pour chacune des 10 prochaines questions, vous entendrez d'abord le numéro de la question; à ce signal, lisez la phrase qui suit ce numéro; choisissez parmi les 3 réponses proposées, celle qui convient le mieux au sens que le mot ou l'expression soulignée dans la phrase initiale confère à cette phrase. Lorsque vous entendrez le numéro de la question qui suit, nous vous recommandons de passer à cette nouvelle question.

Now carefully read item 21.

21. Mary stood up for her friend.

    a) Mary defended her friend.
    b) Mary gave her friend a seat.
    c) Mary rose to talk about her friend.

22. They were playing fairly well.

    a) They were playing about average.
    b) They were playing much better.
    c) They were playing better.

23. When do you eat lunch?

    a) Do you eat lunch at home?
    b) Do you eat lunch in a hurry?
    c) Do you eat lunch at twelve o'clock?

24. <u>Where</u> do you want to go?

    a) Do you want to go to that store?

    b) Do you want to go tomorrow?

    c) Do you want to go by bus?

25. After many years of waiting, he <u>finally</u> went.

    a) In fact, he went.

    b) Finally, he went.

    c) Inevitably, he went.

26. Bob presented the <u>worst</u> report I have ever read.

    a) I have never read such a bad report.

    b) Bob's report is as bad as the one I read.

    c) Bob's report is worse than the one I read.

27. He liked <u>to make up</u> stories to tell his children.

    a) He liked to exaggerate stories.

    b) He liked to tell stories.

    c) He liked to invent stories.

28. That <u>serves</u> you <u>right</u>.

    a) You are getting good service.

    b) You always have a good time.

    c) You deserve what you get.

29. I <u>could have written</u> the letter for her.

    a) Maybe I wrote the letter for her.

    b) I should have written the letter for her.

    c) I would have been able to write the letter.

30. I <u>could have eaten</u> more.

    a) I had eaten enough.

    b) I ate too much.

    c) I was still hungry.

Deuxième section

Directives: Pour chacune des 10 prochaines questions, vous entendrez d'abord le numéro de la question suivi d'une phrase qui vous sera lue à deux reprises; choisissez ensuite, parmi les trois phrases qui vous sont proposées sur le questionnaire, celle dont le sens est le plus rapproché de la phrase que vous aurez entendue.

Now listen carefully to item 31.

31. a) They took the bus fare.
    b) They put the bus fare down.
    c) They put the bus fare up.

32. a) Joan was at home.
    b) Joan was near home.
    c) Joan was almost home.

33. a) Blood is found in the human body.
    b) Blood is formed in the human body.
    c) Blood is necessary in the human body.

34. a) I can't carry it.
    b) I can't accept the pain.
    c) I can't stand the pain.

35. a) How are you getting along in school?
    b) How are you getting away from school?
    c) How are you getting into school?

36. a) What are you examining?
    b) What are you searching for?
    c) What are you finding out?

37. a) Nothing will be done about it.
    b) Nothing will be done in the future.
    c) Nothing will be done now.

38. a) Paul wanted to look at his book.
    b) Paul wanted to find his book.
    c) Paul wanted to see his book.

39. a) I'll discuss it with my parents.
    b) I'll tell it to my parents.
    c) I'll mention it to my parents.

40. a) We won't have any time.
    b) We have plenty of time.
    c) We should hurry.

Première section

Directives: Pour chacune des 20 prochaines questions, vous entendrez d'abord le numéro de la question. A ce signal lisez la phrase qui suit ce numéro; choisissez ensuite, parmi les 3 réponses proposées, celle dont le sens concorde le plus avec la phrase initiale. Lorsque vous entendrez le numéro de la question qui suit, nous vous recommandons de passer à cette nouvelle question.

Now carefully read item 41.

41. It was obvious to me that John knew the answer.
    a) That John knew the answer was obvious.
    b) That I knew the answer was obvious.
    c) John was certain I knew the answer.

42. It was very cold but I walked to the store.
    a) I walked to the store so it was cold.
    b) I walked to the store although it was cold.
    c) I walked to the store because it was cold.

43. Mary has to be driven home.
    a) Mary has to drive home.
    b) Mary has to drive someone home.
    c) Someone has to drive Mary home.

44. Having finished his work, Joe went.
   a) Joe had finished his work.
   b) Joe was finishing his work.
   c) Joe will be finishing with his work soon.

45. He must have been tired playing hockey all day.
   a) Playing hockey must have made him tired.
   b) He was tired of playing hockey.
   c) Playing hockey must be tiring.

46. He ran into the hospital screaming.
   a) Why did he run into the hospital?
   b) How long ago did he run into the hospital?
   c) How did he run into the hospital?

47. Tom had to work to earn money.
   a) When did Tom have to work?
   b) Why did Tom have to work?
   c) Why did Tom have to go to work?

48. Joe screamed with anger.
   a) Why did Joe scream?
   b) When did Joe scream?
   c) How did Joe scream?

49. John will stand by the window.
   a) Where will John stand?
   b) Why will John stand?
   c) How will John stand?

50. While I was having coffee, the doorbell rang.
   a) Why did the doorbell ring?
   b) When did the doorbell ring?
   c) How long ago did the doorbell ring?

51. Come at the same time tomorrow.
    a) How should I come?
    b) When should I come?
    c) For how long should I come?

52. Bob ate breakfast before going to work.
    a) When did Bob eat breakfast?
    b) Can Bob eat breakfast?
    c) Why did Bob eat breakfast?

53. The bird singing in the tree is a robin.
    a) The bird could be singing.
    b) The bird might be singing.
    c) The bird is singing.

54. Allan is so busy that he can't come.
    a) Allan is too busy to come.
    b) Allan is not coming either.
    c) Allan is not coming yet.

55. Mary will be calling us soon.
    a) Mary is going to call.
    b) Mary might be calling.
    c) Mary was going to call.

56. John should have been at the party.
    a) John wasn't at the party.
    b) John went to the party.
    c) John had the party.

57. The baby should be drinking his milk now.
    a) The baby hasn't drunk his milk yet.
    b) The baby has already drunk his milk.
    c) The baby has been drinking his milk.

58. Mark has lived in Montreal for two years.
    a) Mark still lives in Montreal.
    b) Mark lived in Montreal two years ago.
    c) Mark will live in Montreal for two years.

59. Joe will hear Bill talking on the phone.
    a) Bill will be talking on the phone.
    b) Bill must be talking on the phone.
    c) Bill can talk on the phone.

60. Bob shared his lunch with Joe.
    a) Bob will soon share his lunch with Joe.
    b) Bob has already shared his lunch with Joe.
    c) Bob still shares his lunch with Joe.

Deuxième section

Directives: Pour chacune des 20 prochaines questions, vous entendrez d'abord le numéro de la question suivi d'une phrase qui vous sera lue à deux reprises; choisissez ensuite, parmi les 3 phrases qui vous sont proposées sur le questionnaire, celle dont le sens concorde le plus avec la phrase que vous aurez entendue.

Now listen carefully to item 61.

61. a) People believe he will win.
    b) People believe they will win.
    c) He believes he will win.

62. a) Someone asked John to leave the room.
    b) John asked to leave the room.
    c) John asked someone to leave the room.

63. a) Joan must tell Mary this afternoon.
    b) Mary must tell Joan this afternoon.
    c) Mary must be told by this afternoon.

64. a) Roger was driven home.
    b) Roger drove us home.
    c) Roger was driven home by us.

65. a) Tom's meal was served to the waiter.
    b) Tom served a meal.
    c) The waiter served Tom a meal.

66. a) Someone has to ask Joe permission.

   b) Joe has to ask someone permission.

   c) Someone has to ask permission for Joe.

67. a) Someone brought John the book.

   b) Bill brought the book.

   c) John brought the book.

68. a) I didn't want to go.

   b) Mary didn't want to go with me.

   c) I wanted to go too.

69. a) Must john go?

   b) Will John go?

   c) Did john go?

70. a) Bill always wants something.

   b) Bill has always wanted something.

   c) Bill doesn't want anything.

71. a) Won't Bob come?

   b) Didn't Bob come?

   c) Did Bob come?

72. a) Was the work easy?

   b) Is the work easy?

   c) Can the work be easy?

73. a) They are going to phone him.

   b) They were phoning him.

   c) They can phone him.

74. a) The music might be boring.
   b) The music was boring.
   c) The music can be boring.

75. a) Mary shouldn't be hurt.
   b) Mary isn't hurt.
   c) Mary wasn't hurt.

76. a) I must have finished it.
   b) I have finished it.
   c) I haven't finished it.

77. a) They were on the table.
   b) They are on the table.
   c) They must have been on the table.

78. a) Mary isn't a teacher.
   b) Mary must be a teacher.
   c) Mary will be a teacher.

79. a) Joe washed himself.
   b) Joe was washing himself.
   c) Joe washed the boy.

80. a) The meal might be good.
   b) The meal was good.
   c) The meal will be good.

Compréhension orale et écrite.

Première section: Monologue

Directives: Pour chacune des 5 prochaines questions, vous entendrez d'abord le numéro de la question suivi de quelques phrases et d'une question qui vous seront lues une seule fois. Répondez ensuite à cette question en choisissant votre réponse parmi les 3 choix qui vous sont proposés.

Now listen carefully to item 81.

81. a) A club.
    b) A marina.
    c) A gas station.

82. a) Spiders.
    b) Wasps.
    c) Butterflies.

83. a) On a boat.
    b) In a school.
    c) On a farm.

84. a) Taxi drivers.
    b) Policemen.
    c) Criminals.

85. a) At a market.
    b) At a flower show.
    c) At a fashion show.

Compréhension orale et écrite.

Deuxième section: Dialogue

Directives: Avant de répondre aux 5 prochaines questions, vous entendrez d'abord, à deux reprises, un dialogue de quelques répliques. Répondez ensuite aux questions portant sur ce dialogue en choisissant vos réponses parmi celles qui vous sont proposées. Lorsque vous entendrez le numéro de la question qui suit, nous vous recommandons de passer immédiatement à cette nouvelle question.

Now listen carefully to the dialogue.

86. The two people talking are
   a) Unhappy.
   b) Teachers.
   c) Students.

87. The man's problem is that
   a) He needs advice.
   b) His girlfriend doesn't want to see him anymore.
   c) He is missing classes.

88. The man really wants to talk to the woman because
   a) He is thirsty.
   b) He doesn't have a class.
   c) He needs advice.

89. The woman agrees
   a) To talk to the class.
   b) To miss her class.
   c) To talk to Sue.

90. In the conversation, the man is
    a) Unhappy.
    b) Talking to his girlfriend.
    c) Worried about school.

4ième Partie: Compréhension orale et écrite.

Troisième section: Morceau choisi

Directives: Vous disposez de 8 minutes pour compléter la lecture du texte proposé et pour choisir vos réponses aux 10 dernières questions de ce test.

A la fin de la période qui vous est allouée, vous entendrez un signal. Nous vous prions de vous interrompre dès que vous entendrez ce signal, même si vous n'avez pas eu le temps de répondre à toutes les questions.

THE PITCHER PLANT

This plant is called the pitcher plant because of the shape of its leaves. It is in these that insects are trapped and digested by the plant. The secretions of the leaves tempt insects to crawl down them. When the insect has fed itself and is ready to fly away, it finds itself trapped. The tops of the leaves are too narrow to let it spread its wings. The sides of the leaves also hinder its exit. These are lined with little hairs which point downward. When the insect grows tired of clinging to the leaf, it must inevitably fall to the bottom of the plant. This is often partly filled with water. Here, it may starve to death or drown and the plant feeds on the dead body.

91. The leaves of the plant are made in the shape of
    a) Insects.
    b) Little hairs.
    c) Pitchers.

92. The plant feeds upon
    a) Insects.
    b) Water.
    c) Secretions.

93. An insect is attracted to the plant because of
    a) the shape of the leaves.
    b) The little hairs.
    c) The liquid on the leaves.

94. The insect crawls down the leaf in order to
    a) Fly out.
    b) Eat.
    c) Rest.

95. The insect cannot fly out of the leaf because
    a) It is in the water.
    b) The little hairs tickle it.
    c) The top is too narrow.

96. The insect clings to the leaf until it
    a) Dies.
    b) Gets tired.
    c) Escapes.

97. If the bottom of the plant is not filled with water, the insect will
    a) Starve.
    b) Drown.
    c) Eat.

98. Hairs on the sides of the leaves keep this insect from
    a) Getting trapped.
    b) Getting out.
    c) Getting in.

99. Getting caught by a pitcher plant means
    a) An insect will be unable to fly.
    b) An insect will feed itself.
    c) An insect's death.

100.The bottom of the plant often contains
    a) Dead insects.
    b) Water.
    c) Secretions.

Nous vous remercions de votre attention et vous prions de remettre votre questionnaire et votre feuille réponse. Nous vous suggérons cependant de vérifier d'abord si vous avez bien indiqué votre nom, et tous les autres renseignements nécessaires sur votre feuille réponse... Si vous avez des remarques à nous comminiquer, nous vous prions d'en faire part au responsable de l'administration de ce test; ces remarques peuvent nous être utiles.

Merci de votre collaboration.