UNIVERSIDADE FEDERAL DE SANTA CATARINA

PÓS-GRADUAÇÃO EM LETRAS/INGLÊS E LITERATURA CORRESPONDENTE

TEST USEFULNESS IN THE EFL EXTENSION PROGRAM AT UNIVERSIDADE FEDERAL DE SANTA CATARINA (UFSC): AN ANALYSIS

JORGE HUMBERTO SCHADRACK

Dissertação submetida à Universidade Federal de Santa Catarina em cumprimento parcial dos requisitos para obtenção do grau de

MESTRE EM LETRAS

FLORIANÓPOLIS

Junho de 2004

Esta dissertação de Jorge Humberto Schadrack, intitulada TEST USEFULNESS IN THE EFL EXTENSION PROGRAM AT UNIVERSIDADE FEDERAL DE SANTA CATARINA (UFSC): AN ANALYSIS, foi julgada adequada e aprovada e, sua forma final, pelo Programa de Pós-Graduação em Letras/Inglês e Literatura Correspondente, da Universidade Federal de Santa Catarina, para fins de obtenção do grau de

MESTRE EM LETRAS

Área de concentração: Inglês e Literatura Correspondente
Opção: Língua Inglesa e Lingüística Aplicada

_____
Dra. Mailce Borges Mota Fortkamp
Coordenadora

BANCA EXAMINADORA:

_____
Dra. Mailce Borges Mota Fortkamp
Orientadora e Presidente

_____
Dra. Maria da Graça Gomes Paiva
Examinadora

_____
Dra. Adriana Kuerten Dellagnelo
Examinadora

Florianópolis, 25 de junho de 2004.

To Hella Altenburg (in memorian)

ACKNOWLEDGMENTS

Special thanks to:

- My advisor, Dr. Mailce B. M. Fortkamp, for her patience and comfort at difficult times and most of all for helping me become a researcher;
- Vera and André, for the friendship and the accommodation in Florianópolis;
- Ervim, for encouraging me to enter the Master's program.

Other valuable thanks to:

- My family, for supporting me all the time;
- My colleagues and friends at UFSC for both the fun and mutual support;
- Professor Viviane, Adriana, Denise, and all the people at DLLE (I loved working with you!);
- The teachers at DLLE who offered to be the participants in this study!
- To Marta, for being such a good friend!

And finally, thanks to those people and professors at PGI who constantly treat us as human beings, and who made me grow as a person.

Florianópolis, June 2004.

ABSTRACT

TEST USEFULNESS IN THE EFL EXTENSION PROGRAM AT UNIVERSIDADE FEDERAL DE SANTA CATARINA (UFSC): AN ANALYSIS

JORGE HUMBERTO SCHADRACK

UNIVERSIDADE FEDERAL DE SANTA CATARINA
2004

Supervising Professor: Dr. Mailce M.B., Fortkamp

The present study investigated the usefulness of the written tests designed by the teachers of the EFL extension program at Universidade Federal de Santa Catarina (UFSC). The data consisted of one sample of each level's mid-term and final test collected, totaling twenty samples. Each test was analyzed by means of a test usefulness framework proposed by Bachman and Palmer (1996), which consisted of six test qualities, namely *reliability, construct validity, authenticity, instructiveness, practicality,* and *impact.* However, the sixth test quality, *impact,* was not included in the analysis, as it would require specific instruments and extra time to be measured. In addition, a set of interviews with the teachers was carried out in order to substantiate the findings provided by the analysis. The analysis of data revealed that teachers do not base test design on any specific language testing theories, or guidelines. Teachers actually seem to rely on their own intuition and conceptions stemming from their experience in both classroom practice and language assessment. More specifically, with respect to the analysis of usefulness based on Bachman and Palmer's (1996) model, the results showed that none of the tests may be said to contain all usefulness qualities in a balanced fashion. In some tests a quality or two stand out at the expense of another. In other words, among the tests analyzed, different tests contain different usefulness qualities at different extents. It is believed that studies such as the present one may

contribute to a better understanding of the connection between the teaching and testing practices in the present context of research. In addition, suggestions are provided for a special testing training workshop for teachers as a means for treading the path towards more standardized testing practices among teachers in the EFL program at UFSC.

Number of pages: 114

Number of words: 39,116

RESUMO

A UTILIDADE DOS TESTES APLICADOS NO PROGRAMA EXTRACURRICULAR DE INGLÊS COMO LÍNGUA ESTRANGEIRA NA UNIVERSIDADE FEDERAL DE SANTA CATARINA (UFSC): UMA ANÁLISE

JORGE HUMBERTO SCHADRACK

UNIVERSIDADE FEDERAL DE SANTA CATARINA
2004

Professora Orientadora: Dra. Mailce M.B.Fortkamp

No presente estudo investiga-se a utilidade dos testes escritos criados pelos professores do programa extracurricular de inglês como língua estrangeira na Universidade Federal de Santa Catarina (UFSC). Os dados consistiram de uma amostra de cada teste escrito – aplicados na metade e no final de semestre, respectivamente – de todos os níveis existentes, totalizando vinte amostras coletadas. Cada teste foi analisado utilizando-se do modelo de utilidade de teste proposto por Bachman e Palmer (1996), composto de seis qualidades: *confiabilidade*, *validade do construto*, *autenticidade*, *interatividade*, *praticidade,* e *impacto*. Entretanto, a sexta qualidade, *impacto*, não foi incluída na análise, pois esta requereria instrumentos específicos e tempo extra para ser mensurada. Além disso, foram conduzidas entrevistas com os professores a fim de confirmar os resultados obtidos através da análise. A análise dos dados revelou que, ao criarem os testes, os professores parecem não se basear em nenhuma teoria ou orientação na área de avaliação de línguas, confiando em sua própria intuição e concepções adquiridas a partir de sua experiência na prática do ensino e avaliação em sala de aula. Mais especificamente, com relação à analise de utilidade baseada no modelo de Bachman e Palmer (1996), os resultados mostraram que é possível afirmar que nenhum dos testes possui todas as qualidades de utilidade de uma forma

balanceada. Em alguns testes uma ou outra qualidade se destaca em detrimento de outra. Em outras palavras, dentre os testes analisados, diferentes testes possuem diferentes qualidades de utilidades em diferentes níveis. Acredita-se que estudos como este possam contribuir para uma melhor compreensão da ligação entre a prática do ensino e da avaliação no presente contexto de pesquisa. Além disso, propõem-se sugestões para um treinamento em avaliação aos professores sob a forma de oficinas (workshops), a fim de trilhar os caminhos em busca de uma prática mais padronizada de avaliação pelos professores do programa extracurricular de inglês na UFSC.

Número de páginas: 114

Número de palavras: 39.116

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

1.1. Preliminaries

In the second language instruction environment, a great deal of decisions have to be constantly made – whether when placing individuals in a specific level of a course of instruction, when establishing course objectives, choosing the course textbook to be used or planning lessons, among "many other aspects of teaching and learning" (Genesee and Upshur, 1996, p. 3). However, in my own view, taking decisions about individuals is perhaps one of the most crucial moments in the language teaching and learning context. Placing a student in a particular level of a course, or determining whether a student should pass or fail, for instance, calls for a great deal of responsibility from those in charge of taking such decisions. Thus an extremely important and significant, as well as most common method of collecting information in order to make judgments and take decisions concerning individuals is the test (Genesee and Upshur, 1996). It is this high degree of responsibility that is delegated on teachers and educators in general that has motivated me to carry out research specifically on language testing.

In general terms *language testing* plays a powerful role in people's lives, although it has become less impositional and more humanistic over the years (McNamara, 2000). It serves a wide array of purposes as an unquestionable procedure in selection processes - such as those for a job position in a company or agency in which good knowledge of a foreign language is a preliminary condition - for university entry in an English speaking country, for migration purposes and for measuring how much input a learner has achieved at a certain point in an EFL course (McNamara, 2000, pp. 4-5). Regarding the latter, in the mind of some teachers, their pedagogical practice in class may even be influenced by the test that will follow. However, testing should benefit several aspects

of teaching, such as precisely defining weaknesses and difficulties encountered by an individual student or the whole group, evaluating appropriateness of course syllabus, motivating students in their learning by allowing them to show how they perform certain tasks in the language, as well as learn from their weaknesses (Heaton, 1975; 1988, pp. 5-7). Therefore, it is important that language testing and the information it provides be understood by those involved in creating and using tests, in both practical and research contexts (McNamara, 2000, pp. 4-5).

The current literature in language testing - especially in Genesee and Upshur (1996) - often refers to terms such as **assessment** and **evaluation**. In my view these two words are often confused and misunderstood. While *assessment*, commonly referred to as a synonym for the word testing, encompasses the gathering of language information, as well as test information. (Davies, Brown, Elder, Hill, Lumley and McNamara, 1999, p. 11), *evaluation* is the extension of this process for the purpose of making judgments or decisions. (Davies *et al.*, 1999, p. 56). In fact, in a broader and more practical sense, second language evaluation involves mainly decision taking (Genesee and Upshur, 1996), whereas assessment pertains to the instruments used – such as interviews, questionnaires, case studies, and also observation techniques – for the purpose of this decision-taking (Davies *et al.*, 1999). We may thus conclude that the words *assessment* and *testing* are strictly linked.

In the present study I will concentrate on the process of *assessment*, more specifically the design of written language tests. The next section will therefore address definitions of language testing.

1.2. Defining language testing

Generally speaking, language tests measure a person's competence in the first or foreign language. McNamara (2000) defines language testing as

> a procedure for gathering evidence of general or specific language abilities from performance on tasks designed to provide a basis for predictions about an individual's use of those abilities in real world contexts. (p. 11)

In other words, tests should measure a testee' s degree of mastery in the performance of a task that resembles those of real-life. More specifically, Davies *at al.* (1999) describe language tests as "instruments to measure language ability (current capacity to perform an act) and aptitude (potential ability to learn a language)" (p. 1), consisting of "specific tasks through which language abilities are elicited" (p. 107).

Perhaps one of the clearest definitions of a language test is provided by Genesee and Upshur (1996): "a set of tasks requiring observable responses to language or in language that can be scored and interpreted with reference to norms, domains, or instructional objectives" (p. 154). Norms, in this context, are "the descriptions of the performance of clearly identified groups of individuals on the test" (p. 238). A domain is here referred to as a specific area of knowledge or skill, and instructional objectives are the knowledge or skills that a particular lesson, unit, or course contains (Genesee and Upshur, 1996). A task is what a test taker is asked to do during the test, whether it is a "test item involving complex performance in a test of productive skills (speaking and writing)" (Davies *et al.*, 1999, p. 152) or less complex performance, as a "component of language (e.g. grammar item, vocabulary item)" (Davies *et al.*, 1999, p. 92). Genesee and Upshur's (1996) statement about language testing, in my view, is the one that best encompasses the real purpose of assessment. It is therefore the definition I adopt for the present research. Further theoretical aspects regarding language tests will be covered in the next chapter.

1.3. The study

The aim of the present study is to carry out an analysis of the usefulness of the mid-term and end-of-term achievement tests in the EFL extension program at UFSC. The corpus of the proposed analysis consists, therefore, of mid-term and end-of-term achievement tests used in the program, which is run by the University's *Departamento de Língua e Literatura Estrangeiras* (DLLE). For testing purposes, teachers themselves are in charge of both written and oral tests for the levels they teach, based mainly on the

input in the course books used in each level, following their own concepts and criteria, hence making testing in this context a rich source for analysis and research. Test construction does not follow a specific framework, and teachers might be dependent on "misconceptions about the development and use of language tests, and unrealistic expectations" (Bachman & Palmer, 1996, p. 3). The present study consists of evaluating these tests using Bachman and Palmer's (1996) framework of test usefulness, which will be summarized in the next chapter.

Bachman and Palmer's (1996) model of test usefulness seems comprehensive and well grounded theoretically. They argue that all six test qualities, namely *reliability, construct validity, authenticity, interactiveness, impact,* and *practicality,* complement each other and teachers should find appropriate balance among the qualities, depending on each test situation. In other words, deciding on whether there is balance among these qualities in a test it will exclusively depend on a specific test and the specific situation or purpose it has been designed for. In this study, however, due to the scope of the present research, *impact* will not be addressed.

Since the main aim of this study is to carry out a qualitative analysis of the achievement tests of the referred EFL program, using the Bachman and Palmer's (1996) framework, as mentioned before, I pursue the following two central research questions:

1. Do the achievement tests contain all the following usefulness qualities, namely *reliability*, *construct validity*, *authenticity*, *interactiveness*, and *practicality*, as proposed in Bachman and Palmer's (19 96) model?

2. How do teachers design the written tests for the EFL extension program at UFSC?

The first research question has been motivated by Bachman and Palmer's (1996) claim that balance among the usefulness qualities in tests depend on their specific situation and context for which they were designed. The second research question has been motivated by my own interest in the process that teachers follow as well as criteria they adopt when designing the written tests.

For data collection, two samples of each written test for each level of the EFL program were collected (a mid-term test and an end-of-term test), totaling 20 samples of tests. The analysis consisted of investigating the existence of the **usefulness** qualities in the test tasks devised by the EFL program teachers, following suggestions proposed by Bachman & Palmer (1996)

In terms of *reliability*, investigation focused on the criteria the teachers apply when correcting written tests, so that their correction is coherent and uniform. In terms of *construct validity*, the analysis investigated whether the task constructs are related to the syllabus, that is, whether the tasks do evaluate what is intended and reveal students' mastery on specific areas of language abilities defined by the course syllabus.

With reference to *authenticity*, the analysis investigated whether tasks are similar to those suggested in the course book and practiced in class.

Regarding *interactiveness*, my analysis allowed me to look into the degree of test taker involvement with the task. In other words, it allowed me to determine whether or not the test requires that the test-taker apply his or her own topical knowledge while performing the task. An analysis on *practicality* attempted to reveal if the resources required - human resources, material resources and time - are really available, specifically those related to the material used (use of dictionaries, for instance).

Impact will not be addressed, as this test quality is very difficult to measure for the scope of the study. However, a subset of teachers underwent an interview in order to provide more details with respect to how they design tests.

Both deficiencies and positive aspects provided basis for the discussion of results, as well as insights for further research.

1.4. Value of research

Hughes (1989) states that there is no perfect test. A good test will serve the course program, more specifically, it will satisfactorily measure how much students have

achieved in terms of a specific course syllabus or course objectives, as well as promote beneficial *washback*[1] I thus believe that the present study would be the first step towards better testing practice in the context of the present research. More specifically, the outcomes of this study could shed some light upon developing a specific training program in order to establish better understanding of test design and practice among these teachers and their academic coordinators.

In addition, the study might be applied in other EFL instruction contexts. There are other universities around Brazil that run EFL extension programs such as the one at UFSC, and it might be investigated whether it is their teachers themselves who design all written tests, or if the tests are designed by the academic coordinator of the program, for instance. Thus, when applied in further contexts the present study may provide a better understanding of testing practices in EFL extension programs and institutes.

1.5. Organization of the thesis

This thesis is organized in the following way: the first chapter, the introduction, leads the reader into the context of the present research by focusing on the importance and definition of language testing. It also provides an overview of the present study: its main objectives, the research questions, as well its relevance in the context of investigation. Chapter two presents the review of literature in language testing, which includes extended theoretical aspects on language testing, a brief historical background, empirical studies in the area, and the description of Bachman & Palmer's (1996) framework of test usefulness. Chapter three presents the method of data collection, and chapter four describes the analysis in details, as well as the discussion of the outcomes. Finally, chapter fives consists of a summary of results, presenting the limitations of the study, providing insights for further research and bringing pedagogical implications.

---

1 Also known as backwash, it is defined as 'the effect of testing on instruction" (Davies et al., 1999, p. 225), in other words, the consequences that testing brings on the teaching and learning context. Further aspects on washback are discussed in chapter two.

CHAPTER II

GENERAL ISSUES IN LANGUAGE TESTING

The main aim of this chapter is to review relevant literature and studies in the field of language testing. Thereby the following issues will be addressed: first essential theoretical aspects and definitions will be presented; then some historical background on language testing will be overviewed and empirical studies in the area will be discussed. Finally, being the basic framework for carrying out the analysis of the present work, Bachman and Palmer's (1996) test usefulness model will be described.

2.1. Theoretical aspects

In the current literature on language testing scholars highlight the existence of different types of test and test tasks. Genesee and Upshur (1996, p.141) explain that since tests are not a single method of collecting information, different test tasks are basically different ways of eliciting performance from the test taker.

At a first glance, in a macro perspective of the issue, McNamara (2000), for instance, points out the existence of different types of tests, according to their *method* and according to their *purpose*. In terms of method, they can be divided into **paper-and-pencil language tests** and **performance tests**. The former traditionally assess language knowledge components (grammar, vocabulary) or receptive understanding (listening and reading comprehension); the latter assess language as an act of communication, and are traditionally composed of speaking and writing tasks. In terms of purpose, tests can be divided into two main types: **achievement tests** and **proficiency tests**.

Achievement tests aim to measure how much input a learner has accumulated up to a certain moment in a course of study, in order to argue in favor of the preceding teaching practice. Hughes (1989) points out the existence of two types of achievement

tests, *progress* achievement tests, and *final* achievement tests. *Final* achievement tests take place at the end of a course of study, whereas *progress* achievement tests may be administered several times throughout a specific course, measuring how much an individual has learned by then (Hughes, 1989). Further aspects of achievement tests will be discussed in the next section.

Proficiency tests aim at measuring an individual's competence in specific areas of the foreign language for future language use purposes (McNamara, 2000), whether this person has or has not had any training or instruction of the foreign language (Hughes, 1989). This particular competence is also known as **criterion**, which is the candidate's "relevant communicative behavior" (McNamara, 2000, p. 6) in the future 'real life' situation. Proficiency tests may, for instance, test communicative abilities for a specific professional situation (McNamara, 2000), or academic purposes, such as evaluating a potential student or whether he or she is competent enough in a foreign language in order to "follow courses in particular subjects areas" (Hughes, 1989, p. 9). Some proficiency tests have been standardized for worldwide use. Examples of these are the TOEFL, an American examination for candidates who wish to enter American universities, the IELTS tests, for those who wish to pursue university studies in the UK or Australia, among others (Davies, Brown, Elder, Hill, Lumley and McNamara, 1999, p. 54).

Hughes (1989) also mentions the existence of **diagnostic tests** and **placement tests**. Diagnostic tests point out both strengths and weaknesses of students in order to provide basis for the teaching that will follow. However, although they constitute an advantage in more individualized or self-instruction instruction environments, their ideal large size still hinders their practical use. Placement tests help determine what course level a certain student should be placed into, and should be designed by those responsible for the program syllabus (Hughes, 1989).

In terms of approaches to test construction, Hughes (1989) establishes the difference between **direct** and **indirect** testing. Direct testing assesses the candidate's

ability in specific skills, such as speaking and writing. Indirect testing, however, "attempts to measure abilities which *underlie* the skills in which we are interested" (p. 15). In other words, in indirect test tasks the test taker is required to show his or her ability to use the language through tasks (usually written ones) assessing grammatical structures, vocabulary, or even spelling, without performing the real skill in which the use of certain structures or vocabulary is expected (Davies *et al.*, 1999).

Hughes (1989) also contrasts **discrete point** testing and **integrative** testing. In the former, one single grammatical structure is tested individually in one task, while in integrative testing different language elements (such as prepositions, pronouns, verbs, among others) are assessed together in one task. Hughes (1989) adds that these two forms of assessment are related to direct and indirect testing mentioned above: discrete point tests are indirect, while integrative tests are direct.

Two other different kinds of testing that Hughes (1989) mentions refer to scoring of performance. **Norm-referenced** tests yield information about a candidate's performance by comparing it with that of other candidates who took the same test. Conversely, tests that directly provide information regarding what a candidate can do in the language are known as **criterion-referenced** tests.

Two other distinctions in testing are made in terms of scoring. Hughes (1989) explains that testing in which no interpretation by the scorer is needed (such as multiple-choice and gap-filling task tests), are called **objective** testing. However, tests that require a scorer's judgment (such as a composition test) are called **subjective** tests. For Hughes (1989), objective testing is a more reliable means of assessment as there will be no difference in scoring between two different raters in the same test.

In addition, Hughes (1989) highlights one of the most discussed and fairly desired forms of assessment, the so-called **communicative language testing**. This assessment procedure evaluates candidates' performance on real acts of communication, such as the speaking, reading, listening, and writing skills.

As can be observed from the above, there are a number of theoretical aspects that have to be taken into consideration in the study of language testing. However, as the context of the present research is the assessment of student's knowledge in an EFL program I will now concentrate on concepts with respect to **achievement tests**.

2.1.1. Achievement tests

As Davies (1997, in McNamara, 2000) explains, "achievement or attainment tests are concerned with assessing what has been learned of a known syllabus" (p. 87). Genesee and Upshur (1996) refer to achievement tests as *objectives-reference* tests, whose content "is derived from an understanding of the instructional objectives for a particular course, unit, or lesson" (p. 151), so that test tasks should resemble those encountered by students in the classroom practice. Achievement tests may be based either on course objectives or on a course book syllabus (Hughes, 1989).

Weir (1993) states that achievement tests should be closely linked to and reflect the teaching that preceded them. He considers tests as part of the learning process, and therefore importance should be given to students' success, not deficiencies. Heaton (1975; 1988), for instance, adds that achievement tests should be considered tools to encourage good performance in the target language, as well as promote confidence among students. However, it is necessary to make clear what exactly an achievement test is measuring, and how (Weir, 1993).

Achievement test tasks are mainly composed of **test items** (McNamara, 2000) or "parts of a test which require a specified form of **response** from the **test taker**" (Davies *et al.*, 1999, p. 201)**.** McNamara (2000) offers definitions of the most common task items.

A **cloze test**, also known as *gap-filling*, is a test of reading consisting of a text with regularly deleted words which are supposed to be supplied by the test taker (Davies *et al.*, 1999; McNamara, 2000). Heaton (1975; 1988) states that, based on the Gestallt

theory of 'closure', that is, closing gaps in patterns subconsciously, cloze tests "m easure the reader's ability to decode 'interrupted' or 'mutilated' messages by making the most acceptable substitutions from all the contextual clues available" (pp. 16). This type of test item is generally used to measure linguistic abilities according to specific contexts. On average it is used to test vocabulary or some particular grammar point. However, Weir (1993) warns that a few shortcomings should be considered: if the task is not well designed, if the rubrics are not clear, or even if the contextualisation is poor, a candidate might become confused if more than a word is possible in one single gap. Alderson, Clapham & Wall (1995) describe *cloze* items and *gap-filling* items as two distinct types of tasks. In *gap-filling* tasks words or phrases are deleted for the purpose of testing certain linguistic features, unlike *cloze* items, which are more suitable for assessing overall language proficiency by deleting every *nth* word in a written passage.

Short answer questions, a more productive test task, test comprehension (listening or reading) by requiring test takers to show in their own words what they have understood (McNamara, 2000). Besides being a means of avoiding guessing by the candidate, answers may vary from one single word to a couple of sentences, which should ideally be specified in the instructions. In order to ensure reliability, a special marking scheme is needed when judging such answers, including score features, such as spelling and grammar errors (Davies *et al*., 1999), if these are part of the test specifications. Weir (1993) stresses that rubrics should make clear how much of an answer is enough (short factual answers or complete answers) in order to allow adequate judgments when scoring items.

The **multiple choice format** is composed of test items in which candidates have to choose the correct alternative among a number of other optional alternatives (McNamara, 2000). Although it facilitates scoring, its construct validity has been questioned due to the fact that it does not test the production of language, only the ability of the testee to recognize correct forms and reject obviously incorrect options

(Davies *et al.*, 1999, pp. 124-125). Furthermore, it bears little resemblance with real-life language use, does not test language as communication (Heaton, 1975; 1988), and there is the possibility for the test taker to guess the correct answer (Weir, 1993).

Alderson, Clapham & Wall (1995) also describe tasks with **dichotomous items**, commonly known as true/false or yes/no statements. This type of task is easy to mark and is very useful in the assessment of **reading comprehension**. However, as Alderson, Clapham & Wall (1995) warn, a disadvantage of such tasks is the high guessing factor. To overcome this shortcoming, they suggest the inclusion of a third alternative, such as one that reads 'not given' or 'doesn't say'.

Davies *et al.* (1999) provide some more test tasks definitions, such as the **composition test**, "a test of writing in which candidates are required to write one (or more) composition or essay" (p. 27), very common if the EFL program includes the teaching of writing skills, since it often takes "the form of consolidation or extension of the work done in the classroom" (p.136) as well as allows students to show their skills in organizing language material with their own words and ideas, and to communicate (Heaton, 1975; 1988).

Listening comprehension, which tests the candidate's ability to understand spoken language (Davies *et al.*, 1999, p. 110), usually by means of pre-recorded material on tape or CD, might also be tested if the course or program includes it. This material includes dialogs, radio broadcasts, and lectures, among others. Comprehension may be checked via one or more of the tasks defined above. Visual aids, such as pictures, are also commonly used. Regarding the scoring of listening test tasks, Hughes (1989) suggests only comprehension correctness be considered, not grammar or spelling errors (p. 139).

Similarly, a **speaking test**, which is "an assessment of the **ability** to speak the **target language**" (Davies *et al.*, 1999, p. 182), is also usually included in the assessment. Oral test tasks "should elicit behavior which truly represents the candidate's

ability and which can be scored validly and reliably". Tasks include language functi ons, such as *expressing* (thanks, requirements, opinions, apology, among others); *narrating* (a sequence of events); *eliciting* (directions, help, service, clarification, among others); *directing* (ordering, instructing, advising, among others) and *reporting* (description, comment decisions) (Hughes, 1989, p. 101-102).

The above task items are the most common in regular achievement tests. However, Heaton (1975; 1988) adds a few more task items, such as **completion items** and **transformation items**. According to Heaton, completion items are preferred to multiple-choice items since they assess productive language, instead of recognition. In these items, the test taker needs to complete a sentence or question with appropriate words. However, in order to be reliable, completion items should be carefully designed in order not to allow ambiguous interpretation. Transformation items are also more useful than multiple choice ones as they require the test taker to rewrite a sentence in another way, for instance. However, as in the case of completion items, restricting possible answers is very difficult, unless the rubrics are clear enough to avoid that.

Heaton (1975; 1988) also mentions another type of task item called **items involving the change of words**. In this task format, the candidate is supposed to write in the space provided the correct form of an underlined word (which may be a verb in the infinitive, for instance) in a text or sentence. Another type of task item described by Heaton (1975; 1988) is the **broken sentences items**. In this latter type, the candidates are required to write complete sentences out of given "cue words", which may be separated by slashes, a useful grammar or function task item if rubrics are clear and examples are provided.

Pairing and matching items and **combination and addition items** are also mentioned by Heaton (1975; 1988). The former may assess the knowledge of question/response items of a dialog, or even grammatical knowledge, by matching the correct alternatives of two separated columns. Ideally, as Alderson, Clapham & Wall

(1995) suggest, the target (right-hand) column should include additional distracting matching options than the left hand column requires, so that the possibility of default matching is reduced.

As mentioned above, the present research involves the qualitative analysis of achievement tests in an EFL program. Thereby, as will be seen in the analysis of the data collected, the most common test task items that teachers used in the design of the written tests were the following: gap-filling items, short answer questions, multiple choice items, dichotomous items, completion items, transformation items, pairing and matching items, items involving the change of words, broken sentences items, and combination and addition items.

Judgments about the test takers are made based on the results of their performance of the test tasks. The next section thus addresses important issues regarding test scores and rating procedures.

2.1.2. Scores and rating procedures

Test results are commonly referred to as **scores**, the ultimate data used for making decisions about individuals (Bachman and Palmer, 1996, p. 194). Test scores are the representation of the test taker's individual attributes or characteristics (Genesee and Upshur, 1996).

Bachman and Palmer (1996) define **score** in two ways. One refers to 'the number of test tasks successfully completed" (p. 226) allowing the sum of correct responses. Thus, criteria for 'successful completion' and right or wrong responses need to be defined by the test designer. The other refers to the development of scales for language ability, in which both the components of a test's construct as well as its performance levels should be defined. Levels should range from the lowest (meaning 'no evidence of' the ability measured), to the highest (meaning 'evidence of mastery of' the ability measured) (p. 226-227). In the scoring process of tests, the performance of test takers

has to be judged and graded in order to provide valid and reliable results, obtained following rating procedures.

Rating differs from scoring in that it represents the quality of language performance in a composition or in an oral task, while scoring represents the number of correct answers in tasks that do not require rater interpretation (Genesee and Upshur, 1996). McNamara (2000) defines **rating procedure** as the "agreed procedure followed by **raters** in judging the quality of performances, particularly in the assessment of speaking and writing" (p. 136). Despite the slight difference, ratings may be converted into scores in order to be part of a test grade (Davies *et al.*, 1999). A good example of a rating procedure would be the correction and scoring of an essay. The rater (usually a teacher) may arrive at a single score based on his or her judgment of a test taker's writing ability, taking into consideration the candidate's use of appropriate vocabulary, textual organization, and register, among others.

Both scoring and rating procedures are prone to yielded drawbacks. Scores may be affected by the nature of tasks (Genesee and Upshur, 1996), that is, they require certain skills from the test taker that are independent from the test content itself. Among some of the factors that lead to these shortcomings, it has been observed that a test taker's experience on a specific test task favors his or her final score on that task. In sum, a test taker performance and score in a test may be a reflection of not only the assessed content, but also of the nature and format of the tasks (Genesee and Upshur, 1996).

Research has also shown that **rating**, although necessary, may be problematic. Teachers, especially in the assessment of the speaking skill (and also the writing skill, as exemplified above), are tempted to directly provide a score based on a "single impression of the impact of the performance as a whole", a procedure called **holistic rating** (McNamara, 2000, p. 43). McNamara (2000) explains that ratings are subjective, that is, the rating given to a test-taker reflects not only his or her performance in a

certain task, but also the level of personal judgment made by the rater. What is more, ratings may seriously vary according to different raters, as well as the different occasions of the performance. In order to reduce these differences, McNamara points out the need for the establishment of rating criteria, so that a "basic framework or orientation for the rating process" (p. 38) may be determined. He thereby suggests the design of analytic rating, specifically in the case of oral assessment, in which different aspects of the performance (fluency, appropriateness, pronunciation, grammar, vocabulary, among others) are analyzed separately, under a pre-established detailed **rating scale**.

Also known as *proficiency scale*, a **rating scale** is made of a series of constructed levels against which aspects of a testee's oral or written performance are judged. In oral performance, for instance, this scale ranges from *zero* mastery to an end-point (which represents *native-like* performance), and is made of levels or *bands* (which characterize a testee's proficiency or ability in a certain performed task) (Davies *et al*., 1999, p. 153-154). Likewise, when marking compositions (assessing a testee's writing skills), the following performance aspects can be included in the rating scale: content, organization, cohesion, and vocabulary (Alderson, Clapham, & Wall, 1995, p. 107-109). Genesee and Upshur (1996) also advocate the use of rating scales as they yield more reliable information and feedback, which ensues teacher and student reflection over both oral and written performance before and after tests. If there is the need to arrive at a single score, it could be obtained by adding the sub-scores determined by the level or band of each performance aspect of a rating scale.

In order to improve the fairness of rating procedures, McNamara (2000) suggest that raters undergo special training, in which level descriptors and rating categories are discussed and agreed by raters together in meetings. McNamara adds that rating differences may not be eliminated completely, but after some of these training sessions and with constant monitoring of rater performance, these differences will be reduced

significantly. Alderson, Clapham & Wall (1995), for instance, emphasize the need for the calculation "intra-rater reliability" and "inter-rater reliability". **Intra-rater reliability** refers to the degree of similarity in judgment of one rater over one composition or oral performance on two different occasions; whereas **inter-rater reliability** pertains to how similar two different raters judge the same composition or oral performance.

It is probably safe to assume that expecting exactly the same marks in either means of rater reliability would be unrealistic, but Alderson, Clapham & Wall (1995) state that, specially in the case of oral and written examinations boards of important English language proficiency tests, some of their pre-established standards have to be met.

Let us now draw attention to what has been taking place in terms of research in language testing. The following section will thus address empirical studies and major findings in the field.

2.2. Language testing research and practice over time

In an extensive article reviewing modern language testing, Bachman (2000)[2] explains that language testing research and practice has been subject to extensive study and refinement since the 1980s. While in the sixties and the seventies the main concern in second language testing focused on the four skills and their components (grammar, vocabulary, pronunciation), in the eighties, influenced by the work of researchers such as Henry Widowson (1978; 1979; 1983), Sandra Savignon (1972; 1983), Michael Canale and Merril Swain (1981), and Keith Morrow (1979), language use seized to be considered an "isolated trait" and began to be viewed as the "creation of discourse, or

[2] Bachman's (2000) article is, in my view, a milestone in the language-testing field. It is thus used as a spine for the historical background in this review of literature. I have made efforts to refer to the original sources cited in the article, but only a few have been found. Therefore, in order not to exclude relevant information, most sources referred in the article are cited in the present review as "in Bachman (2000)".

the situated negotiation of meaning, and of language ability as multicomponential and dynamic" (Bachman, 2000, p. 3). In other words, language testers began to take into consideration a whole discoursal and sociolinguistic aspect of language use, and began to apply a communicative approach to language assessment, which caused language testing to establish its position of relevance in the field of applied linguistics by the end of the eighties. In fact, in an earlier article (Bachman, 1991), Bachman supports that language testing should be considered a "discipline in its own right", an argument that had just been presented by scholars such as Alderson (1991), Bachman (1990a), and Skehan (1988,1989,1991) (Bachmann, 1991, p. 671). Bachman (2000) also states that there were two other landmarks in the eighties and nineties. The first refers to the use of language tests as research instruments in second language proficiency acquisition (with studies by Allen, Cummins, Mougeon, and Swain, 1983; Harley, Allen, Cummins, and Swain, 1987; 1990). The second refers to Pienemann, Johnson, and Brindley's (1998) research paper on test design and scoring based on the sequence in which the language learner develops his or her proficiency (Bachman, 2000).

Proficiency representation for language assessment, however, seems to be the focus of extensive debate. Chalhoub-Deville (1997), for instance, reviews "models of proficiency that have influenced second language testing in the past twenty five years" (p. 3). She explains that, unfortunately, among scholars there is no single agreed model for representation of proficiency. What researchers do is adapt from different proficiency models, which generally lack empirical bases and do not provide clear directions when language assessments must be designed, nor "contribute to the lack of congruence between theories and test construction" (Alderson, 1991, in Chalhoub-Deville, 1997, p. 4-10). However, she holds that while some existing theoretical models have their limitations, other meaningful and useful, empirically based assessment frameworks, such as Hinofotis, Bailey and Stern's (1981) and Chalhoub-Deville's (1995a), proved to be a powerful and valid tool in assessing language proficiency as

they were designed to be used in their specific contexts. Hinofotis, Bailey and Stern's (1981) model, for instance, was developed for the purpose of assessing L2 oral proficiency of foreign teaching assistance, while Chalhoub-Deville's (1995a; 1995b) model was created in order to establish the components employed by native speakers of Arabic on three oral tasks (an interview, a narration, and a read-aloud task) designed to assess the proficiency of intermediate-level learners of that language (Chalhoub-Deville, 1997).

Qualitative research, as Bachman (2000) states, has also been subject to increasing interest as the focus shifts to the test taker's performance, characteristics, processes, and strategies in test tasks, as well as language testing discourse. Bachman points out that several qualitative research approaches, including expert judgments, verbal reports, questionnaires, interviews, text analysis, conversational analysis and discourse analysis, should be a valuable tool for research refinement. In addition, as Bachman holds, attempt should continue to be made in combining both qualitative and quantitative approaches, something that has already been done in studies by Anderson, Bachman, Cohen and Perkins (1991), Weigle (1994), Clapham (1996), North (1996), and Sazaki (1996).

Practical advances have also taken place. Bachman (2000) highlights the creation of *testing cross-cultural pragmatics*, developed at the University of Hawaii (Hudson, T., Detner, E., & Brown, J.D. 1992; 1995, in Bachman, 2000), which consists of assessment instruments in order to obtain data on both cross-cultural pragmatics and pragmatic competence in cross-cultural communication. He also points out the improvement of *language testing for specific purposes* (Alderson and Clapham, 1993; Clapham, 1996, among others, all cited in Bachman, 2000), which has broaden its use in several branches of science and technology areas.

Among factors that affect performance on language tests, Bachman (2000) explains how research has been carried out involving characteristics of the testing

procedure, the test-taking process, and characteristics of test takers. In terms of *characteristics of the testing procedure* studies ranged from test item characteristics and difficulty (e.g. Anderson *et al.*, 1991; Freedle and Kostin, 1993; Perkins and Brutten, 1993; Perkins *et al.*, 1995; Bachman *et al.*, 1996; Fortus *et al.*, 1998; Freedle and Kostin, 1999, all cited in Bachman, 2000), performance of different task types (e.g. Riley and Lee, 1996; Fulcher, 1996; McNamara and Lumley, 1997, among others, all cited in Bachman, 2000), and differences in rating behavior (e.g. Brown, 1995; Chalhoub-Deville, 1996; Milanovic *et al.*, 1996, among others, all cited in Bachman, 2000). The interest in the *test-taking process* itself and test taker strategies, as Bachman continues, has made way for studies by Buck (1991); Wijgh (1996); Storey (1997), among others. Van Lier (1989); Berwick and Ross (1996), and Lazaraton (1996) have carried out research in oral interview discourse, and Wigglesworth (1997; 1998) has investigated 'the effects of planning on test performance" (Bachman, 2000, p. 11). Research in *characteristics of test takers*, such as academic background, native language, culture, gender, field dependence, occupation, aptitude, background knowledge, and personal characteristics, includes studies by Hill (1993), Sasaki (1996); Sparks *et al.* (1998); Clapham (1993; 1996); and Berry (1993).

In Brazil, specifically, debates take place concerning new tendencies and paradigms in evaluation in a broader educational context, as alternatives to traditional assessments. Paiva (2000), for instance, argues in favor of taking into consideration test takers' personal characteristics in order to propose a more holistic and humanistic form of assessment. Supporting Buttler's (1995-6) concepts of learning styles, Paiva (2000) believes that assessment and evaluation should take place by taking into consideration the different ways students feel and process what has been taught in class. Paiva (2000) admits that doubts, anxiety and resistance still constitute significant obstacles towards the use of new forms of assessment and evaluation, but, on the other hand, Buttler's

(1995-6) psychoanalytic and holistic view of the human being shed light into the search for a better teacher-student relationship in terms of evaluation.

Another alternative to traditional forms of assessment is what scholars call *authentic* or *performance assessments*, which consist of specific authentic tasks (i.e., tasks eliciting real-life performance) assessing a candidate's abilities in specific study or professional situations (Davies *et al.*, 1999). Authentic assessments have received new attention due to recent advances in teaching methods and educational measurement procedures, which have led to a better understanding of test task design and usefulness (Bachman, 2000). The search for the development of standard-based assessment and the increasing criticism towards standard multiple-choice test caused researchers, such as Herman *et al.* (1992); Wiggins, 1989; 1993); Newman *et al.* (1998); Twillinger (1997; 1998); Aschbacher (1991); Schavelson *et al.* (1992); Swanson *et al.* (1995); Solano-Flores and Shavelson (1997), all cited in Bachman (2000), to support more authentic assessment, based on performance, specially in the field of language teaching (e.g. Harrison, 1991a; 1991b; Rea Dickins, 1991a; 1991b; Kohonen, 1997; Brown and Hudson, 1998, all cited in Bachman 2000), thus keeping close the communicative language type of testing. However, despite this increasing awareness regarding authentic assessment, Bachman (2000) warns that further research in this field is still needed. Further aspects regarding alternative and performance assessment are discussed below, in the section dealing with empirical studies.

One may now wonder about what the future will hold in terms of language testing. Bachman (2000) believes that language testing and testers must grow in terms of professionalization and validation research. It is also urged that the profissionalization should also focus on the training of language testers, as well as 'the development of standards of practice and mechanisms for their implementation and enforcement" (Davies, 1996, in Bachman, 2000, p. 19). Bachman (2000) adds that our ability to look into the validity of our test scores and interpretations, as well as the fair

use of these score judgments, will be benefited with resources to be revealed in years to come.

Expressing a more skeptical view of current language testing, McNamara (2000) points out the crisis in the area, "masked by impressive appearance of technological advance, such as computer based testing" (p. 79). Bachman (2000), however, remains optimistic about the future, explaining that research has shortened the distance between language testers and applied linguists, and advances in testing technology have brought sophistication and more variety in test formats and procedures. Scholars such as Green, (1988), Ginther and Chawla (1997) - in Bachman, 2000 - for instance, suggest that these new task formats allow new insights in **validity** investigation and redefinition of test constructs and scoring, in contrast with those of paper and pencil task formats. Nevertheless, in order to explore the potential of these new technologies, a good amount of collaborative work among language testers and researchers is obviously still needed (Dunkel, 1996, in Bachman, 2000).

Finally, Bachman (2000) summarizes that "our long history of validation research" has largely benefited studies in language testing (p. 24). The effect of factors and processes on language performance is better understood, and so is the use of an array of research tools and their positive and negative aspects. In addition, debates over test ethics have become more consistent, and test construct validation has been combined with test use consequences.

After having addressed language testing under a historical perspective, the next section will deal with empirical studies, as well as further discussions regarding washback, and ethics and morality in language testing.

2.3. Empirical studies in language testing

The practical advances previously discussed above are strictly connected to empirical studies in several areas of language testing. To the best of my knowledge,

much of what has been discussed in terms of achievement tests is extensively discussed in specific literature. Heaton (1975; 1988), Hughes (1989), Weir (1993), and Alderson, Clapham, & Wall (1995), for instance, provide extensive guidance on how to produce, administer, and score language tests to be used in the L2 instruction environment, whereas Genesee & Upshur (1996) look into language assessment under a wider perspective, discussing evaluation as a means for improving both teaching and learning. Besides dealing with the same aspects of the field as the formers do, McNamara (2000) presents a critical analysis of both traditional and newer tendencies in the language testing practice. Bachman & Palmer (1996) address language testing in a more sophisticated and elaborative manner by presenting a framework for test usefulness, which is the model adopted for the present study.

Unfortunately published empirical studies in the specific area of *achievement testing* have not been found. I will therefore discuss those studies and aspects that helped me build up a better understanding of the main issues debated by language testing scholars.

Research on vocabulary testing, for instance, has investigated several tests, such as the TOEFL vocabulary test items (Schmitt, 1999), a word association test for measuring L2 proficiency (Wolter, 2002), Nation's (1983; 1990) Vocabulary Levels Test (Laufer and Nation, 1999), Nation's (1990) revised versions of the 2000 Word Levels and University Word Level Vocabulary Tests (Beglar and Hunt, 1999).

Schmitt's (1999) study investigated what vocabulary items in the TOEFL test measure, more specifically, what kind of world knowledge is elicited from the candidate, and what is known by the candidate about the tested items. Using thirty L2 learners of English as participants, Schmitt's (1999) study has shown that a correct answer in the test was not an indicator that the testee knew all meanings or collocations of a specific word. It has also turned out that the testee's inferenc ing skills, rather than the knowledge of a word, did also influence the choice of a correct answer.

Wolter (2002) carried out a study that included thirty participants, investigating the development of both a multiple word association test and a second test for measuring L2 proficiency. Just as in past studies, the outcomes of Wolter's (2002) study failed to show conclusive evidence on why L2 proficiency can be successfully assessed by means of a word association test, and one of the main shortcomings was the deviation in the scoring interpretations.

Laufer and Nation (1999) developed a study in order to investigate the validity of Nation's (1983; 1990) Vocabulary Levels Test, as this test format is strictly related to language competence (Grabe, 1991; Frederiksen, 1982, in Laufer and Nation, 1999), and is also a helpful tool in placement tests of EFL programs. In their study four groups of foreign language learners at different proficiency levels were used as participants. Using complex statistical methods of measurements, Laufer and Nation (1999) brought evidence that the tests are reliable in measuring a testee's vocabulary growth. Similarly Beglar and Hunt (1999) investigated the reliability and validity of the revised versions of the 2000 Word Level and the University Word Level Vocabulary Tests. In this latter study, also using statistical measurement, both the 2000 Word Level and the University Word Level Vocabulary Tests were found to satisfy the minimal demands. Taken together these studies show that, although a great deal of vocabulary assessment may successfully measure the **size** of vocabulary knowledge (i.e. how many words are known), measuring the **depth** of vocabulary knowledge of individuals is still a crucial issue that warrants further research (Schmitt, 1999).

Alternative language assessment is another important issue that has been the focus of discussions. Aiming at helping teachers in deciding the type of language test they can use depending on the context of instruction and purpose, Brown and Hudson (1998) have developed a list of alternative assessment characteristics, which stemmed from the combination of previous lists by Aschenbacher (1991), Herman, Ashenbacher, and

Winters (1992), and Huerta-Macías (1995). In this list Brown and Hudson (1998) cite that alternative assessments (p. 654):

1. require students to perform, create, produce, or do something;
2. use real-world contexts or simulations;
3. are nonintrusive in that they extend the day-to-day classroom activities;
4. allow students to be assessed on what they normally do in class every day;
5. use tasks that represent meaningful instructional activities;
6. focus on processes as well as products;
7. tap into higher level thinking and problem-solving skills;
8. provide information about both the strengths and weaknesses of students;
9. are multiculturally sensitive when properly administered;
10. ensure that people, not machines, do the scoring, using human judgment;
11. encourage open disclosure of standards and rating criteria; and
12. call upon teachers to perform new instructional and assessment roles.

Brown and Hudson (1998) enforce that although alternative assessment is an exciting and tempting procedure, its reliability and validity must be measured carefully. They argue that just like other assessment methods,

> the designers and users of alternative assessments must make every effort to structure the ways they design, pilot, analyze, and revise the procedures so the reliability and validity of the procedures can be studied, demonstrated, and improved. (p. 656)

Hamp-Lyons (1997), for instance, contends that alternative assessment should be given the same importance as traditional assessments, as it cannot be assumed that it will bring positive washback, and has thereby devised a framework of classroom/learner performance behaviors. This framework, which stemmed from Meisels, Dorfman and Steele's (1995) model of learner performance characteristics in performance assessment and standardized tests, has aided the researcher in the investigation of "whether actual behaviors do, in fact, show the features predicted by the model" (Hamp-Lyons, 1997, p. 301).

Two other issues to be addressed in language testing research are authenticity, which, more specifically, pertains to what extent a test task may be considered

authentic, and performance assessment. Some test tasks are more authentic than others, depending on the relation of these tasks and the test taker's target language use (TLU) tasks (Bachman and Palmer, 1996). Lewkowicz (1997; 1999, in Bachman, 2000) carried out two studies to investigate test takers' perception of authenticity in test tasks. The results showed that test takers' judgments were highly influenced by their performance and familiarity with the tasks. Lewkowicz (1997, in Bachman, 2000), however, states that authenticity in language tests still calls for further research.

Performance assessment is carried out requiring candidates to perform tasks that replicate performance in real-life contexts (Davies *et al.*, 1999). Among studies regarding performance assessment that have emerged in the nineties, Bachman (2000) cites those by Dunbar, Koretz and Hoover (1991), Linn, Baker and Dunbar (1991), Mehrens (1992); Moss (1992), Baker, O'Neil and Linn (1993), among others, who have worked on applying issues such as **reliability**, **validity**, and **impact** in performance assessment. Camp (1993); Condon and Hamp-Lyons (1993); Hamp-Lyons (1996); Lynch and Mc Namara (1998), among others, all cited in Bachman (2000), have looked into other specific types of performance assessment, such as writing portfolios and oral interviews. Research has also been conducted involving aspects of oral interviews both quantitatively (Elder, 1993; Bachman, Lynch and Mason, 1995; Brown, 1995; Lumley and McNamara, 1995; among others, all cited in Bachman, 2000); and qualitatively (Ross and Berwick, 1992; Young, 1995; Lazaraton, 1996; Kormos, 1999; among others, all cited in Bachman, 2000). Language performance assessment stemming from task-based language teaching has also been widely discussed by Norris, Brown, Hudson and Yoshioka (1998), who suggested ways for designing both language performance language teaching and testing (Bachman, 2000). In fact, Bachman (2000) points out the increasing design of textbooks which include language performance tasks, such as portfolios, conferences, journals, among others, which help language teachers and test designers better understand the nature of performance assessment and its usefulness.

Washback effect as well as ethics and morality in language testing have also been subject to continuing debating, though more extensive empirical research, specifically in the case of washback, is still needed (Bailey, 1996). Washback, in the case of both achievement and proficiency tests, can be characterized as the effect of a test on the teaching that takes place before or after it (Buck, 1988, in Bailey, 1996), whether positive (or beneficial), which "occurs when assessment procedures correspond to the course goals and objectives (Brown and Hudson, 1998, p. 688), or negative, which is characterized by the lack of relation between the goals and objectives of a course curriculum and the testing procedures it adopts (Brown and Hudson, 1998). Scholars agree that tests affect and influence (at different extents) all those involved in the pedagogy context, such as students, teachers, course books, course content and method, and coordinators, among others. (Hughes, 1993, and Alderson and Wall, 1993, both cited in Bailey, 1996).

Bailey (1996) states that the scarcity of empirical research on washback effect is due to the difficulty in providing measurable variables, as well as the extremely close relation of washback and other teaching and learning variables. Bailey (1996) rounds up by stating that beneficial washback from tests can only be obtained under the following conditions: all those involved with test must fully understand the purpose of the test; score reports should be clear, detailed, and informative; test takers must be able to "find the results credible and fair" (Bailey, 1996, p. 275); if students' performance is to be measured by an external-to-program test, the latter should be related to the target language program's curriculum. In addition, tests should be based on updated accepted theoretical principles, they should preferably contain authentic texts and tasks, and the test takers should be "invested in the assessment process" (Bailey, 1996, p. 277).

Concerns with the **ethics** and profissionalization of language testing have also aroused during the past twenty years. Several scholars, such as Stanfield (1993), Davidson, Turner and Huhta (1997); Hamp-Lyons (1997a), and Norton (1997), all cited

in Bachman (2000), have shown their concern with issues strictly related to ethics and morality in language testing: the social, political, and educational consequences of the use of language tests, as Davies (1997) explains:

> While the growing profissionalization of language testing is perceived as a strength and a major contribution towards a growing sense of ethicality, the increase in commercial and market forces, as well as the widespread use of language assessment as an instrument in government policy, may pressure language testers into dangerous (and unethical) conduct (p. 236).

It is common sense among scholars that the main aim of ethics is to balance "the individual and the social" (Davies, 1997, p. 237), to maintain social justice without undermining individual differences. Ethics is also strictly connected to morality, and most times the two are used interchangeably (Davies, 1997). As Davies (1997) posits, morality pertains to "codes of practice", and "constrains responsibility within reasonable limits" (p. 238).

Spolsky (1981; 1997), Bachman (1990), Shohamy (1993a; 1993b; 1997), and Lynch (1997), all cited in Bachman (2000), argued that language tests may be harmful if not used correctly and ethically, addressing the theory that language tests may abusively serve gatekeeping, political, and educational purposes. Shohamy, Donitsa-Schmidt and Ferman (1996), for instance, argue that:

> Results obtained from tests can have serious consequences for individuals as well as for programmes, since many crucial decisions are made on the basis of test results. The power and authority of tests enable policy-makers to use them as effective tools for controlling educational systems and prescribing the behavior of those who are affected by their results – administrators, teachers and students (p. 299).

Spolsky (1997) reports that gatekeeping tests have been used for control and power for thousands of years, and expresses his concern with the gatekeeping function of tests, whose main function is "to determine qualifications for positions or for training for positions" (p. 242). Given the limitations that still persist in psychometrics, Spolsky (1997) warns us that we must reflect about the confidence placed in examination results that lead to decision-making about people, as well as be careful in order not to justify

our judgments claiming that tests are simply "more than a lottery with a bias in favour of those who do better on it" (Spolsky, 1997, p. 246). However, as the researcher states, these uncertainties are being accepted as being inevitable, and at best testers might as well use what he calls multiple testing and alternative methods, and interpret outcomes carefully and cautiously. Decisions after gatekeeping tests must be based on human judgment, and not on software or mechanical test marking processes (Spolsky, 1997, p. 246).

Despite all efforts to make tests ethic, Davies (1997a) warns "it is not possible for a tester as a member of a profession to take account of all possible social consequences" (p. 335). Davies explains that if a test is used for a purpose other than that to which it has been designed, than its designer cannot be blamed for any eventual undesirable consequence. The lack of a proper code of ethics and its sanctions for the event of incorrect conduct, the scarceness of subjects volunteering for research, the lack of moral or ethical conduct that some researchers may adopt, all these factors lead to the necessity of an "'ethical milieu' through education" (Hofman, 1991, in Davies, 1997a, p. 336). Thus, the International Language Testing Association (ILTA) has devised a 'Code of practice for foreign/second language testing', for test designers and test users. This milieu, however, would be an institutionalized association, with regularized membership, office and officers, and also publications. Members would need licensed qualifications for their professional practice, based on standards and behavior (Davies, 1997a, p. 337). In my personal view, even in more restricted contexts, such as in the case of achievement evaluation in EFL programs, it is possible to develop an internal code of practice. Such an endeavour would eventually help establish the first steps towards a standardized design of useful written and oral tests, which is extremely desirable.

Having addressed important theoretical and historical aspects in language testing, as well as the main areas in which empirical research and debates on language testing

have taken place, we shall now concentrate on Bachman and Palmer's (1996) framework of test usefulness, the main instrument of the data analysis inside the present research.

2.4. Bachman and Palmer's (1996) framework of test usefulness

Bachman and Palmer (1996) define test usefulness as "the essential basis for quality control throughout the entire test development process" (p. 17). They consider test **usefulness** the most important quality of test design and development, and thus propose a model of test usefulness, which includes "six test qualities – reliability, construct validity, authenticity, interactiveness, impact and practicality" (p. 17).

Hughes (1989, in Bachman and Palmer, 1996) contends that although test qualities are sometimes in conflict, there is no reason for totally abandoning any. Bachman and Palmer (1996) state that it is necessary to try to find a balance among them, and that this balance will be different among different testing situations. In sum, a test can only be considered useful if all six qualities are combined.

Bachman and Palmer (1996) also advocate that a useful test must be developed considering its real purpose, the group of individuals who are going to sit it, and the specific language use domain - the Target Language Use (TLU) domain. Test design and evaluation is extremely subjective, that is, it "involves value judgments on the part of the test developer" (Bachman and Palmer, 1996, p. 19).

According to Bachman and Palmer (1996), there is a main difference in **purpose** regarding both the teaching materials in a language program and the test. While the former's pivotal aim is promoting learning, the latter's main aim is to measure how much content has been learned. Bachman and Palmer, however, state that four of the test qualities are strictly related to the learning aspect: the *authenticity* of a language sample practiced in class, the *interactiveness* of a learning task performed in class, the *impact* of a certain activity practiced in class, or the *practicality* of a certain teaching

approach. The two remaining test qualities, *reliability* and *validity*, on the other hand, are very related to the testing or "measuring" situation, as they are strictly related to scores and ratings, which yield numbers used as basis for making judgments and decisions (Bachman and Palmer, 1996). Below, I have summarized each of the test qualities described by Bachman & Palmer in their model:

Reliability: Reliability is defined as "the consistency of measurement" (p. 19). In other words, a reliable test will offer the same score results, whether a particular test taker sits the test on one occasion and setting or another. An example of reliability would be if a certain composition would receive the same score, regardless of the rater who scored it. Reliability is important in the way that it minimizes variations between the tasks used in the test and the classroom tasks in which the target language is used for teaching and learning purposes: the TLU (Target Language Use) tasks. In addition, it provides evidence of how successful test designers have been in minimizing these variations.

Construct validity: Construct validity "pertains to the meaningfulness and appropriateness of the *interpretations* that we make on the basis of test scores" (p. 21). In order to justify a certain test score interpretation, we need logical evidence, which will depend on the test's *authenticity* and *interactiveness*, described below. Construct validity provides evidence that what is intended to be measured is really measured.

Authenticity: Authenticity refers to the level of proximity the test task has to the Target Language Use (TLU) domain task. The TLU domain refers to specific language use tasks encountered in contexts other than the test itself. The TLU domain referred to in this research context is that of the L2 classroom, referred to as the language instruction domain. An example of authenticity would be an oral test question regarding an issue that has been widely discussed in class.

Interactiveness: This quality refers to "the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task" (p. 25). These

characteristics are the student's "language ability (language knowledge and strategic competence, or metacognitive strategies), topical knowledge, and affective schemata" (p.25). An interactive test task would be one in which the test-taker has the possibility to relate the test task's input topical content to his or her own topical knowledge. A test task in which there is a text about raising children, for instance, will be much more interactive if the test-taker is a mother.

Impact: Here this test quality will be restricted to a *micro level*, more specifically how students are affected by sitting a test. Bachman and Palmer state that specific values and goals will affect our using of tests and "our choice will have specific consequences for, or impact on, both the individuals and the system involved" (p. 30). For test takers, the impact will be on taking the test and preparing for it, the feedback on their performance in it and the decisions made about them based on their test scores. For teachers, on the other hand, impact will be on teaching. A negative effect would be what Bachman & Palmer call "teaching to the test" (p. 33), in which the instructional program is adapted to the test, and not vice-versa, very common in cases in which the test has a problem of *authenticity*.

Practicality: Bachman and Palmer define it as the relationship between resources (human resources, material resources, and time) required from the moment a test is designed to the moment it is administered and scored, and those resources available for test administration per se. If a test use requires only resources that are really available, then a test is said to be practical.

In the present study the usefulness quality *impact* will not be used in the analysis, as this quality is very difficult to measure for the scope of the study. In the next chapter I will describe the study procedures (method).

CHAPTER III

METHOD

3.1. The context of research

As previously stated, the data of the present study consists of the mid-term and end-of-term achievement tests used in the EFL extension program at UFSC (Universidade Federal de Santa Catarina), as well as of the interviews with the teachers who designed these tests. The EFL extension program is run by the *Departamento de Língua e Literatura Estrangeiras* (DLLE) of UFSC (Universidade Federal de Santa Catarina), which also offers German, French, Spanish and Italian courses. The courses offered in the English program, also called "extracurricular" (since they are not part of the university's regular syllabus/curriculum in Foreign Languages), were created to offer university students the opportunity to study a foreign language at more reasonable price without compromising the quality. The course books adopted in the EFL extension program are the series "New Interchange" and "Passages" series, both published by Cambridge University Press, and the English teachers are mainly post-graduate students of the university's masters and doctoral program, both in English language. All English teachers are selected after a micro-teaching session, and previous experience in teaching a foreign language is a preferred requisite.

The EFL program is composed mainly of under-graduates (young adults), although these courses are all open to the community. The EFL program is divided into the following levels: levels 1, 2 and 3 (**basic**); levels 4, 5 and 6 (**pre-intermediate**); levels 7 and 8 (**intermediate**); levels 9 and 10 (**advanced**). Each course level is divided into one "term" of 60 class/hours. Classes are usually of 90 minutes, twice a week, although there are groups that attend one 180-minute class once a week.

There are three tests during the semester: one written test is administered in the middle of the term and another in the end of the term, and teachers also administer an oral test at the end of the semester.

3.2. Data collection

The above-mentioned *Departamento de Língua e Literatura Estrangeiras* (DLLE), which runs the EFL extension program at UFSC, is supposed to file all tests devised by the teachers throughout the semesters. Thus it is the teachers' duty to provide photocopies of both mid-term and final tests they design for each level, each semester. Thus, the first step was to examine these files and collect a pair of tests (mi-term and final tests) for each of the above-mentioned levels, totaling 20 samples of tests. Since the EFL program consists of several groups of the same level, taught by different teachers, the following criteria had to be established: each pair of test collected should have been designed by the same teacher in the same semester from 2001 up to 2002. For level one, for instance, the written tests collected consisted of the mid-term and final term test designed by the same teacher, in the second semester of 2002. Recent tests were preferred since both "New Interchange" and "Passages" series have been adopted since 2001. Although these files were made available, a few problems arose.

The files, consisting of plastic envelopes labeled according to the course levels, should contain several complete pairs of tests, of the same level, with at least one pair designed by the same teacher. However, the files contained scattered samples of mid-term and final tests. Moreover, most tests did not have identification, neither mentioning the name of the teacher who designed them, nor mentioning the semester and the year they were applied. Table 1 provides a clear picture of what has been encountered in each envelope.

Table 1: The EFL extension program test files (identification features).

Level	Mid-term test			Final test			Complete pair (mid-term + final tests, by the same teacher)	
	Quantity	Name of teacher?	Semester + Year?	Quantity	Name of teacher?	Semester + Year?	Quantity	Semester + Year?
Level 1	1	Yes	No	2	Yes	No	1 pair	Yes
	1	No	No					
Level 2	2	No	No	1	Yes	Yes	0	
				1	Yes	No		
				1	No	No		
Level 3	1	Yes	Yes	2	No	No	2 pairs	No
	1	Yes	No				1 pair	Yes
	1	No	No					
Level 4	2	No	No	3	Yes	Yes	0	
				1	Yes	No		
Level 5	3	Yes	No	2	Yes	Yes	1 pair	Yes
	1	No	No	1	Yes	No		
				1	No	No		
Level 6	1	No	No	3	Yes	No	1 pair	Yes
							2 pair	No
Level 7	3	Yes	Yes	2	Yes	Yes	2 pair	No
				1	Yes	No		
Level 8	2	No	No	1	Yes	Yes	1 pair	No
				2	Yes	No		
				2	No	No		
Adv.1	1	Yes	Yes	0			0	
Adv.2	1	Yes	Yes	1	Yes	No	1 pair	No
TOTAL	**21**			**27**			**12 pairs**	

The table shows the difficulty encountered in order to pursue the first means for data collection. First, all tests were classified according to their level. Next tests were separated according to their stage (mid-term or final test) and identification characteristics: name of the teacher who designed them and the date, the semester and year they were administered. Mid-term and final tests designed by the same teacher in the same semester were identified as "complete pairs".

For level 1, the following mid-term tests were found: one test identified by the name of the teacher only, and another with neither piece of information (name of

teacher or semester). Two final tests were found, and both included only the name of the teacher. Only one complete pair was found.

For level 2, two mid-term tests were found, but neither had teacher nor semester identification. Three final tests were found, but either they did not include the name of the teacher, or the semester. No complete pair was found for level 2.

For level 3 three mid-term tests were found: on contained both the name of the teacher and the semester, another contained only the name of the teacher, and a third that did not contain neither piece of information. Two final tests were found with neither piece of information. Three complete pairs were found, but only one contained the information regarding the semester.

For Level 4 there were only two mid-term tests and neither contained any information regarding the teacher or semester. Three final tests contained both name of teacher and semester, and one contained only the name of teacher. No complete pairs were found for that level.

For level 5 four mid-term tests were found: two contained only the name of the teacher, and one did not contain either the name of the teacher or the semester. Of the final tests, two contained both pieces of information, one contained the name of the teacher only, and a third did not contain either piece of information. On complete pair was found containing the information regarding the semester.

Level six envelope contained one mid-term test with neither the name of the teacher nor the semester. Three final tests were found, and all three did not contain the information regarding the semester. Three complete pairs were found: one containing the semester, and two others without this piece of information.

For level seven, three mid-term tests were found and all three were identified with the names of the teachers and the semesters. Three final tests were found: two contained both the name of the teachers and the semesters, and one that contained the name of the

teacher only. Two complete pairs were found, but neither includes information regarding the semesters.

For level 8, two mid-term tests with neither pieces of information (teachers or semesters) were found. Five final tests were found, one containing both name of teacher and semester, two containing only the name of the teacher, and two others containing neither the name of teacher, nor the semester. One complete pair was found, but it did not include the semester.

The advanced 1 envelope contained only one mid-term test with both the name of the teacher and the semester, whereas advanced 2 contained the following: one mid-term test with both the name of the teacher and semester, one final test with the name of the teacher only, and one complete pair without the information about the semester.

Of a total of 72 tests, there were thus 21 mid-term tests, 27 final tests, and 12 complete pairs. The figures on the table show that only a minority of teachers provides photocopies of the tests they design to the DLLE (Departamento de Língua e Literatura Estrangeiras).

The test files have revealed another interesting finding: 22 tests (out of a total of 72) contained or were completely made of tasks that had obviously been cut and pasted from the course book teacher's guides. I have referred to these teacher's guides in order to confirm this finding. In addition, most of the tests in the files were designed through the years of 1997 and 1999. Only a few updated tests were found, but the teachers who designed these tests were no longer part of the EFL program staff.

Since no recent complete pairs of tests were found, it was decided that another procedure for data collection needed to be adopted. In other words, new samples of written tests designed during the most recent semesters (the year 2002) had to be collected directly from teachers. Two main factors affected this decision: First, the files were incomplete and not updated; second, this contact with the teachers would provide

the possibility to arrange a further meeting with some of them for the purpose of an interview, in order to provide further evidence regarding the design of written tests.

In order to contact the extra curricular teachers, a list of teachers was provided by the DLLE (Departamento de Língua e Literatura Estrangeiras), containing information, such as the levels they taught in the previous semester, their e-mail addresses and telephone numbers. As I myself had been one of the teachers of the EFL extension program, more specifically for level 8, I decided that my tests should be analyzed as well. For the present study the test samples for level 8 were therefore those I had designed myself.

For the collection of the other test samples, some of the teachers were contacted by telephone; others had to be reached via e-mail. While a few volunteered promptly and either handed me the test samples personally, or sent them to me as attached files through e-mail, others were not immediately willing to reply to my messages, which sometimes had to be sent more than twice.

One of the greatest difficulties lied in the fact that there were not as many teachers who taught upper levels of English as there were for basic and pre-intermediate levels. For level seven, for instance, of two teachers to be contacted only one agreed to participate as a subject. In addition, a few samples of tests were incomplete, so these teachers had to be contacted more than once in order to clarify a few doubts, or provide any extra material that was missing. In both mid-term tests for levels advanced 1 and 2, which were designed by the same teacher, for instance, she was not able to provide the original photocopied articles which had been transformed into a listening comprehension exercises). Due to the above-mentioned difficulties the period of test sample collection lasted five months.

Once volunteer teachers were contacted and the twenty samples of tests were collected, each test was carefully analyzed under the following procedure: first tasks were identified in terms of the constructs to be assessed - namely reading and listening

comprehension, grammar and functions, vocabulary, and writing skills – which are the language abilities taught and practiced in both course book series: "New Interchange" and "Passages". Next, the text books (namely the student's book, the work book, and teacher's guide) were referred to for two main reasons: the first reason was to identify what exactly each test task assessed, more specifically in terms of grammar and functions, and vocabulary content, as well as topics of listening and reading comprehension tasks, and writing skills. This allowed me to also investigate whether there was uniformity in the assessment of these contents (for instance, whether the teacher excluded any grammar point or vocabulary of any specific unit of the course book, or included any grammar point or function that was not supposed to be assessed) The second reason was to investigate the extent to which each test task resembled those tasks observed in the course book (in both the student's book and the work book), and also whether the task rubrics (the instructions for each task) were consistent in order not to allow misinterpretation by the testee. A third step consisted of counting the number of task items in order to investigate whether there was uniformity in the number of items per test task, and finally, the scoring system (provided the teacher included on in his or her test) was analyzed as to investigate whether or not all tasks in a particular test were equally weighted in terms of scoring.

After having analyzed each test under the aspects described above, the tests were then investigated in terms of their usefulness using five qualities of Bachman and Palmer's (1996) framework of test usefulness, namely *reliability*, *construct validity*, *authenticity*, *interactiveness*, and *practicality.* As mentioned in the previous chapter, one of the qualities, *impact,* has not been used in this study, as the measuring of this quality alone would require further instruments, more elaborate procedures and extended research time. The next section will deal with the interviews with teachers, which took place a few months after the test samples were collected.

3.3. The interview with teachers

As I previously stated, a second step in the present study consisted of interviews with the teachers, which have been carried out in order to complement the findings and provide a better understanding of the teachers' criteria and assumptions when designing the tests.

The interviews consisted of pre-established questions used to trigger the discussions and obtain the teachers' views. The questionnaire, composed of eight open questions, addressed specific stages in the testing process, from test design through scoring, including details such as how teachers decided on test content (grammar, functions, and vocabulary), topics, as well as specific skills to be assessed (reading comprehension, listening comprehension, and writing or compositional skills), task formats, test length, and the scoring system (see appendix B).

The first question elicits how the teacher decide about the written test content, more specifically in terms of grammar and functions, and vocabulary. The second question addresses what specific skills the teacher intends to assess in the written test, such as reading, listening and writing. The third question aims to investigate how the teacher chooses the topic and the material for the reading and listening comprehension tasks, and the composition. The fourth question pertains to the task formats chosen for the tests, whether the teacher creates them or collects them from a different source. The fifth question aims to elicit the teacher's concept of the length or size of a written test, and the sixth question addresses how the teacher prepares the scoring system of the tests, which tasks are worth more than others and why. In the seventh question the teacher is asked to provide more specific details regarding how different types of tasks are scored, contrasting gap-filling tasks with those in which the testee is required to write a subjective answer, and also how compositions or essays are scored.

In the last question (question eight) the teacher is given the opportunity to express his or her opinion about the course book used, or about the EFL extracurricular course at UFSC itself.

Originally, as I myself was one of the teachers whose tests were used in the present study (more specifically the tests for level 8) all other teachers were contacted in order to make appointments for the interviews, totaling seven teachers. However, as one of the teachers did not wish to be interviewed, unfortunately only six out of seven teachers volunteered to participate. Four interviews were recorded and later transcribed, and as two of the teachers were not available for a personal contact, their interviews were carried out via electronic mail. The complete set of interviews may be referred to in appendix C. Each teacher is identified with letters, such as teacher A, teacher B, and so forth.

Transcription procedures were adapted from those used in Fortkamp's (2000) study in the area of speech production. Pauses the teacher made in order to think are indicated in parentheses, as in "*(pauses to think)"*. Sound stretches are indicated by colons (:), as in "I:", and filled nonlexical pauses are indicated by 'uh", 'uhm", and 'uh-uh". A period indicates falling intonation and a question mark indicates rising intonation. An exclamation mark indicates that the teacher expressed enthusiasm.

The next chapter thus addresses details regarding the analysis of test usefulness, the teachers' views through the interviews and the discussion of results obtained through the analysis.

CHAPTER IV

THE ANALYSIS OF USEFULNESS OF THE WRITTEN TESTS APPLIED IN THE EFL EXTENSION PROGRAM AT UFSC.

This chapter presents the analysis of test usefulness of the written tests applied by eight teachers in the EFL extension program at Universidade Federal de Santa Catarina (UFSC). The first section provides information about the Target Language Use (TLU) domain, a concept proposed by Bachman and Palmer (1996) to refer to the context in which the target language (in this case English) is practiced in non-test situations. In other words, the first section will depict the environment in which the target language is taught and practiced (more specifically with respect to the textbook used in class) and to understand the relationship between the instructional context and testing practice at the above referred EFL extension program.

The following section consists of the analysis of the above mentioned written tests, and thus the following steps are taken: first each test is described and analyzed in terms of construct and content assessed, task characteristics (task rubrics and format), and scoring system. Second, the written tests are analyzed in terms of usefulness using Bachman and Palmer's (1996) framework of test usefulness. The section that follows deals with the interviews with the teachers who designed the written tests used for the present analysis, in order to confirm hypotheses drawn from the results of the usefulness analysis of the written tests. The teachers' views are discussed following specific stages in test design and scoring. In the last section both the results of the usefulness analysis of the written tests and the teachers' views revealed through the interviews are discussed, and finally the research questions are addressed and answered.

4.1.The Target Language Use (TLU) domain

Bachman and Palmer (1996) argue that the usefulness of a language test lies in the demonstrable correspondence between the test tasks, as well as the test takers' response to them, and the language use in the Target Language Use (TLU) domain.

As Bachman and Palmer (1996) state, the TLU domain comprises tasks of language use practiced by the testee outside the test situation, that is, the "situations in which language is used for the purpose of teaching and learning of language" (p.44), referred to as the language instruction domain. Given the nature of the object of analysis (the EFL extension program achievement tests at Universidade Federal de Santa Catarina), the context in which the target language is being used is thus where English is taught and practiced: inside the classroom. As Bachman and Palmer (1996) warn, in a context where it is difficult to determine the students' real-life domain, that is, the real use of the target language outside the classroom, it is preferred to design test tasks that resemble those of the instructional context. Thus the level of testee performance that teachers aim to observe through the language abilities assessed in the test will reflect the performance of the same language abilities performed in the classroom. In addition, if the test tasks are familiar to the testees, this will improve and optimize their test performance. (Bachman and Palmer, 1996).

The course of each semester (level) in this researched context is determined by the course book syllabus used in class. Thereby, we may be led to conclude that the content of the written tests should be based on the syllabus. Hughes (1989) refers to this approach as the "syllabus-content approach", and that this constitutes an appealing means for the design of a fair test (Hughes, 1989, p. 11). Hughes (1989), however, holds that in order to design a fair test, its content should be based on pre-established course objectives, regardless of the course book syllabus, since "it will provide more accurate information about individual and group achievement, and it is likely to promote a more beneficial backwash effect on teaching" (Hughes, 1989, p. 11). In addition, test content based on badly designed course book syllabi, will inevitably provide inaccurate and

misleading results. In other words, good scores in tests will not indicate that individuals have achieved the course objectives (Hughes, 1989). Nevertheless, in the context of the present study, it is the syllabus of the course book adopted by and used in the EFL program that determines the target language use (TLU) domain (the context in which the target language is used in non-test situations), establishes the course objectives, and, consequently, the content of the achievement written tests.

In the present study it is not my intention to discuss or judge the EFL extension program course objectives per se, or the course book syllabus adopted, but some important issues will be discussed further, in the conclusion section, more specifically regarding the washback effect of the achievement written tests to be analyzed.

At the time of the data collection for the present study, the EFL extension program at UFSC adopted two course book series: from Level 1 through Level 6 the course book adopted is the three-volume New Interchange series, by Jack C. Richards, with Jonathan Hull and Susan Proctor as co-writers, published in 1997 by Cambridge University Press. From Level 7 through Advanced 2, the course book adopted is the two-volume Passages series, by Jack Richards and Chuck Sandy, also published by Cambridge University Press, in 1999. Although it is not the main aim to judge or qualify the nature of the material adopted, it is relevant to briefly describe the course books' tasks, so that the instruction domain is well understood.

The main course components of the New Interchange series are the student's book, the workbook, the class CD audio program, and the teacher's guide. Each of the three volumes (books) of the New Interchange series consists of 16 six-page units, and because each semester of the EFL extension program covers eight of the course book units, each volume of the series is used for two semesters of the EFL extension program, that is, levels one and two use the first volume of the series – the first eight units of the first volume are dealt with in level one, while the last eight units of the first volume are dealt with in level 2. Levels three and four use the second volume of the series – the first eight are dealt with in level three, while the last eight units are dealt

with in level four. Levels five and six use the third volume of the series – the first eight units are dealt with in level five, whereas the last eight units are dealt with in level six.

According to the author of the series, Richards (1997), each unit is divided into two topical/functional "cycles", which follow about the same task/exercise order. In the first "cycle" a topic i s introduced through a short oral activity based on updated real-world cross-cultural information. Then, an audio-recorded conversation (dialog) introduces the new grammar and function of the unit in a situational context. Next, grammar (and parallel functions) are explained in summary boxes and then practiced via accuracy-controlled written tasks and freer communicative oral tasks. The second "cycle" contains a second dialog, which introduces the second grammar point and function of the unit. In addition, in either "cycle", new vocabulary is taught productively (through written and spoken tasks) and receptively (through reading and listening tasks). Each volume contains a review unit after every 4 units. As far as the four skills are concerned, each unit contains a listening task, a speaking task, a reading task, and a writing task, which are strictly linked to the unit's content and topic. The series' workbook, which may or not be used in class, provides extended written practice on each unit's grammar, flu ency and vocabulary, including writing and a reading skill tasks (Richards, 1997).

The Passages series, which consist of two volumes, resembles the New Interchange series in almost every aspect. Jack Richards again is the main author, with Chuck Sandy as co-writer. According to the authors, this multi-skills series (speaking, listening, reading, and writing) serves as a sequel to the New Interchange series and thereby takes students from upper intermediate to advance stages, comprising the four last semesters of the EFL extension program: Level 7 to Advanced 2, respectively: The first six units of Passages One are dealt with in level 7, while the last six units are dealt with in level 8. The first six units of Passages Two are dealt with in level Advanced 1, while the last six units are dealt with in level Advanced 2.

The components of the series are the student's book, the workbook, the audio program CDs, and the teacher's guide. According to Richards and Sandy (1999), each of the two volumes (books) consists of 12 eight-page units divided into two thematic lessons: lesson A and lesson B. There is a review unit after every 3 units. In lesson A, oral or aural fluency activities introduce the unit's topic based on updated real-world cross-cultural information, followed by a grammar summary box with controlled (grammar exercises) and less controlled accuracy practice (pair and group discussion, or a listening exercise). The last activity of the lesson, a writing activity, provides composition skill practice. Lesson B follows the same order and approach of lesson A, but provides a reading exercise, instead of the writing, and a vocabulary exercise. All written tasks, as well as the specific four-skill exercises, are related to the lesson's specific topic, and authors adopt almost an identical approach to that of New Interchange, described above. The series also has a workbook, used either in class, or as homework, providing extra written practice on each unit's grammar, fluency, vocabulary, as well as writing and reading comprehension skills (Richards and Sandy, 1999).

For the purpose of this analysis, some of the most common written tasks (grammar, vocabulary, listening, reading, and writing) in both the New Interchange and Passages series' course book and workbook, will be briefly listed below. Despite the expected increase in level of complexity and difficulty throughout each course level's syllabus, the basic nature and format of written tasks remains the same. In other words, as both course book series have been written by the same author, it has been observed that the type of tasks in both series are almost identical. For grammar and function practice the following task types have been observed to be the most common:

Gap-filling: A task containing dialogs, paragraphs, or isolated statements/questions with gaps in which the students have to insert a word or two from a given list, or even the correct from of a word in parentheses.

Short questionnaire: a task in which questions have to be answered, either with a simple short answer, or a complete answer, according to the rubric. Tasks in which the students have to write a suitable question to a given answer, have also been observed.

Matching columns: a task in which the students have to match two columns, usually pairing questions and answers, choosing the correct response to what has been said, matching clauses to make logical statements, among other possibilities.

Multiple choice: a task in which students have to choose the correct word, phrase, or response.

Sentence completion: a task in which students have to complete sentences, either just by writing about their personal experience, by transforming prompt words, or by giving information about a prompt picture.

Sentence transformation: tasks in which students have to either rewrite sentences using prompt words or expressions in parentheses, or produce one single sentence by combining two other given sentences.

Word addition: a task containing isolated sentences or a paragraph in which some words belonging to a specific word class, for instance, are missing. The students need to recognize the spot, and place these words where they are missing.

For vocabulary practice the following task types have been observed to be the most common:

Word maps: a task in which students have to complete spaces of specific lexical areas with words or phrases taken from a list provided, as shown in the example below (taken from New Interchange One, student's book, p. 8):

Complete the word map with jobs from the list.

architect
receptionist
company director
flight attendant
supervisor
engineer
salesperson
(among others)

PROFESSIONALS		SERVICE OCCUPATIONS
. . . . .		. . . . .
. . . . .		. . . . .
. . . . .		. . . . .
	JOBS	
MANAGEMENT POSITIONS		OFFICE WORK
. . . . .		. . . . .
. . . . .		. . . . .
. . . . .		. . . . .

Picture labeling: a simple task in which students have to label prompt pictures (objects, situations, among others) with words or phrases taken from a list provided.

Chart completion: a task in which spaces in a table have to be completed with words or phrases taken from a list provided. This is a common task for word collocation practice, and lexical grouping, as shown in the example below (taken from New Interchange One, student's book, p. 42).

Find other two words or phrases from the list that are usually paired with each verb.
an art exhibition - a vacation - a party - a trip – shopping – a lot of fun - the dishes – dancing - a play - the laundry

did	housework	. . . . .	. . . . .
went	swimming	. . . . .	. . . . .
had	a good time	. . . . .	. . . . .
saw	a movie	. . . . .	. . . . .
took	a day off	. . . . .	. . . . .

Matching columns: a task in which the students have to match words and their definition, words and their opposites, among other possibilities.

Circling the odd word out: a task in which students have to recognize, from a group of four words, the one that does not belong in that specific lexical group.

Connotative classification: A task in which the students classify words (adjectives, nouns, adverbs, among others) according to their positive or negative connotation.

Multiple choice: a task in which the students have to choose the correct word or phrase. The number of distractors may vary from task to task.

Gap-filling: A task containing dialogs, paragraphs, or isolated contextual statements/questions with gaps in which the students have to insert a word from a list provided.

For reading comprehension practice the following task types have been observed to be the most common:

Checking the correct boxes: a task in which students are required to check the boxes that refer to the correct information according to the text read.

True-false statements: a task in which students are required to check whether statements are true or false.

Chart completion: a task in which spaces in a chart have to be completed with notes based on information extracted or inferred from the text.

Multiple choice: a task in which the students have to choose the correct answer for a question. The number of distractors may vary from task to task.

Short questionnaire: a task in which questions have to be answered, either with a simple short answer, or a complete answer, according to the information obtained from the text.

Inferring word meaning: a task in which the students have to infer the meaning or definition of a few underlined words in the text.

For listening comprehension practice the following task types have been observed to be the most common:

Chart completion: similar to those in reading comprehension practice, this is a task which requires students to complete spaces in a chart by taking notes on information extracted or inferred from the audio recorded passage, usually a conversation.

Answering questions: Specifically in the New Interchange series, this task is very common after each of the unit's printed conversations. In this task the students have to take notes in answer to one or three simple comprehension questions by listening to a second, non-printed part of that conversation

Checking the correct boxes: a task in which students are required to check the boxes that refer to the correct information obtained or inferred by listening to short audio passages.

Unlike the groups of tasks discussed above, which tend to be of the same type or format throughout all levels of both course book series, the writing tasks differ from the most basic to the advanced levels. Besides reinforcing each unit's topics and grammar points, the writing tasks become more complex throughout the course levels as students' compositional skills improve. In the New Interchange series, which covers the first six semesters of the program, the tasks are more practical (writing a postcard, descriptions, narratives, reviews, among others) and usually require students to write no more than two short paragraphs. For each task the course book provides an example, or a model to be followed, usually the first three or four sentences of a paragraph.

In the Passages course book series, which covers the four last semesters of the program, the writing tasks become more complex and concentrate on more specific steps for compositional writing. As Richards and Sandy (1999) discuss, the first volume of the series focuses on "using topic sentences, identifying the main ideas and supporting details, and organizing paragraphs" (Passages 1, student's book, pp. iv and v). The second volume continues this process, focusing on "various genres, such as book reports, comparison and contrast, summaries, business letters, and personal experiences" (Passages 2, student's book, p. iv). In each task of the series, before producing their own compositions, students work on a model text, looking into topic

sentences and paragraphs, among other important features. These writing tasks require students to write more than three paragraphs. The number of words is not clearly specified in neither series' task directions, but as Richards, Hull and Proctor (1997) suggest, it is expected that teachers follow the writing process steps, such as writing first drafts, revising, and editing.

More complex tasks can be found in the workbook and require more attention from the student, as they are usually based on pictures, graphs, tables, or other prompts. Some of these complex tasks may consist of two other integrated tasks: a simple one, which could be any of those discussed above (choosing a correct word, matching two columns, among others), followed by a task that requires better thinking, either word/sentence completion, or sentence writing. It has been observed that most of these tasks were especially designed to promote the practice of very specific grammar points, functions, or vocabulary items related to certain units. In addition, these may, or may not call for the student's own experience and opinion.

4.2. The analysis of the written tests

The main focus of this section is the analysis of usefulness of the written tests. Therefore the following few steps are followed: first, both tests of each course level are briefly described in terms of task characteristics, scoring procedures and possible shortcomings. Secondly, the tests are analysed in terms of usefulness by means of Bachman & Palmer's (1996) framework. Thirdly, the interview with teachers is addressed and the teacher's views regarding test design are exposed. Finally, the results of the analysis and information revealed by the interviews are discussed together.

In the following section each test is briefly described in terms of task characteristics, scoring information, and assessed content.

4.2.1. The description of the tests

One of the primary aspects to be revealed in the description of written tests that consist the data for the present study is uniformity regarding the tests' construct and content assessed, task characteristics, and, if the test provides it, the scoring system (distribution of points among tasks). More specifically, each test was analyzed by observing the following: whether all components or constructs proposed by the text book and practiced in class are assessed (namely listening and reading comprehension, writing skills, as well as grammar and functions, and vocabulary knowledge); whether the test content is assessed with uniformity[3], whether test tasks resemble those in the course book, whether task rubrics are clear enough and contain examples in order to avoid misinterpretation by the testee; and finally, regarding scoring procedures, whether there are tasks that are weighted in terms of scoring (that is, if there are any tasks that are worth more than others).

Additionally, each test analyzed has generated a table that depicts each test's task in terms of its construct focus, the number of items, and the course book content or topics it covers. These tables represent a useful tool for understanding what exactly each task assesses.

Level 1: mid-term (New Interchange 1, units 1 through 4): The mid-term test sample selected for the analysis in the present study is composed of twelve tasks assessing reading and listening comprehension, grammar and functions, vocabulary, and writing skills (see appendix A). Four tasks (tasks seven, eight, nine, and eleven) were extracted from the teacher's guide test; two of them contained slight changes in their items so that they do not look like exact copies of that test. Among the remaining eight tasks, seven tasks, although they have been slightly adapted, resemble those tasks found in the

[3] Uniformity in the present study is characterized by the following condition: when all test tasks contain the same or almost the same number of items, thus equally assessing the contents and constructs proposed in the test.

course book (see appendix A). Task twelve, the paragraph writing task, is the only one that has not been observed in the first four units of the course book. Table 2 presents information regarding each task in the mid-term sample test designed for level 1.

Table 2: Level 1, mid-term test (New Interchange 1, units 1 through 4): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Reading comprehension	3 items	Exchanging personal information
Task 2	Listening comprehension	2 items	Exchanging personal Information
Task 3	Grammar and functions	13 items	Unit 1, cycles A + B
Task 4	Grammar and functions	5 items	Unit 1, cycle B
Task 5	Grammar and functions	4 items	Unit 2, cycle A; Unit 4, cycle B
Task 6	Grammar and functions	5 items	Unit 2, cycle A
Task 7	Grammar and functions	4 items	Unit 3, cycles A + B
Task 8	Grammar and functions	5 items	Unit 3, cycle B
Task 9	Vocabulary	3 items	Unit 2
Task 10	Grammar and functions	3 items	Unit 4, cycle B
Task 11	Grammar and functions	3 items	Unit 4, cycle A
Task 12	Writing	Paragraph	Giving personal information (units 1 to 4)

As can be observed from table 2 above, tasks one and two assess reading and listening comprehension respectively, whose main topic is *exchanging personal information*. Task twelve assesses writing skills and its topic is *giving personal information*. Tasks one, two, three, four, five, six, seven, eight, ten, and eleven focus on the grammar and function content of units 1 to 4 of the course book. Task nine focuses on the vocabulary content of unit 2 of the course book.

Table 2 also reveals that the grammar and function, and vocabulary content of units 1 to 4 of the course book is not assessed with uniformity in this test. In other words, the whole grammar content in unit 1 of the course book, for example (*Wh or Yes/No questions and statements with the verb to be in the Present Simple*), is assessed in two different tasks, namely tasks three and four. The grammar content of unit 2, cycle A (*Simple Present Wh- questions and statements)* is assessed via seven items, in two different tasks (task five and six), while the grammar content of unit 2, cycle B (*time expressions)* is not assessed at all. The vocabulary task (task nine) assesses the knowledge of unit 2 only, which consists of words related to *work* and *workplaces*.

Examining the test rubrics also reveals other possible sources of misunderstanding (see appendix A for the mid-term test for level 1). In six out of twelve tasks the rubrics may cause misinterpretation by the testee. In the first and second tasks (reading and listening comprehension), for example, whose rubrics are "read the text and answer the questions" and "listen to the conversation and answer the questions", respectively, it is not specified whether the testee should answer the questions with complete answers or just factual answers. The testee may not know how much information he or she is supposed to give as an answer. The same problem might arise in the third task, which reads "complete the conversations": the rubrics neither specify the number of words, nor give additional information regarding what to write in the gaps (such as the verb tense). In the writing task (task twelve) the rubrics are: "Write a paragraph about yourself. Use the information you studied from units 1 to 4 (at least five lines)". The task requires the testee to use information covered in the four units of the course book, but the rubrics do not specify what kind of information to be included in the paragraph, nor do they provide any guideline questions that could aid in the building of the paragraph. Specifying the type of information to be elicited in each task is important in order to avoid misinterpretation by the testee. However, the main shortcoming is the fact that paragraph writing is very limited in the first four units of the course book used for level one of the EFL extracurricular program at UFSC. The result is that different students might produce paragraphs of different number of words with different amounts of information, undermining desirable rating and scoring procedures by the teacher. We may thus conclude that the main problems of this test are the lack of uniformity of the content assessed and task rubrics that lack extended directions.

Regarding scoring, the absence of scoring or marking procedures written on the test hinders further assumptions about score distribution among the test tasks.

Level 1: final test (New Interchange 1, units 5 through 8): This final test sample selected for the analysis in the present study, composed of eleven tasks, is a literal

photocopy of test two taken from the first volume of the New Interchange teacher's guide (see appendix A). Instead of numbers, all tasks are identified by letters. The test assesses listening and reading comprehension skills, grammar and functions, vocabulary, and writing skills. Ten out of eleven tasks resemble those observed in the course book. Table 3 provides details each task in this final test sample.

Table 3: Level 1, final test (New Interchange 1, units 5 through 8): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task A	Listening comprehension	4 items	Topics from units 5 through 8
Task B	Grammar and functions	4 items	Unit 5, cycle A
Task C	Grammar and functions	4 items	Unit 5, cycle B
Task D	Grammar and functions	4 items	Unit 6, cycle A
Task E	Grammar and functions	5 items	Unit 6, cycle A + B
Task F	Grammar and functions	5 items	Unit 7, cycle A
Task G	Grammar and functions	5 items	Unit 7, cycles A + B
Task H	Grammar and functions	5 items	Unit 8, cycle A
Task I	Vocabulary	6 items	Unit 8
Task J	Writing	Paragraph	Unit 8
Task K	Reading comprehension	4 items	Unit 8

Table 3 shows that task A assesses listening comprehension, whose topics are those of units 5 through 8. Tasks B through H focus on the grammar and functions content of units 5 through 8 in the course book. Task I, J, and K, respectively, focus on the vocabulary content, writing skills, and reading comprehension of unit 8 in the course book.

The table reveals that the grammar and function content covered in each unit's cycles in this test, namely through tasks B through H, is assessed in a more uniform manner than it is in the mi-term test. However, it may also be noticed that the vocabulary of units 5, 6, and 7 is not assessed. Task I, the vocabulary task, assesses only the vocabulary covered in unit 8.

In this sample of the final test task the lack of clear instructions may hinder the testee performance (see appendix A). Task B, for instance, whose rubrics are "Complete each conversation. Use the present continuous (for example, *is going, are taking*)", unlike similar tasks found in the course book used for level 1, does not provide the base-

form verbs in parentheses. In task C, whose rubrics are "rewrite these sentences using determiners", these determiners to be used in the sentences could have been listed. The rubrics in task E, "Read each conversation and complete the question", could be clearer if they mentioned that the verb tense to be assessed is the Present Simple. Moreover, the only tasks to provide examples in their rubrics are tasks C and H. The main shortcomings of this sample of the final test for level 1 are the lack of uniformity in the assessment of vocabulary, and the lack of extended directions in four out of eleven tasks.

Regarding scoring, during personal communication, the teacher who used this final test revealed that for this specific test she referred to the score system in the first volume of the New Interchange teacher's guide, which suggests that tasks A, B, C, D, J, and K be worth eight points each, tasks, E, F, G, and H be worth ten points each, and task I be worth twelve points. The total score of this final test is thus one hundred.

Level 2: mid-term test (New Interchange 1, units 9 through 12): The mid-term test sample selected for the present analysis consists of six tasks that assess grammar and functions, and writing (see appendix A) of units 9 through 12 of the New Interchange One course book. Tasks one, two, three, and five resemble those observed in the course book. However, task six, a writing task, does not resemble any of the tasks observed in the course book. Table 4 presents information regarding the tasks in this mid-term test designed for level 2.

Table 4: Level 2, mid-term test (New Interchange 1, units 9 through 12): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Grammar and functions	Not mentioned	Unit 9, cycle A
Task 2	Grammar and functions	5 items	Unit 9, cycle A
Task 3	Grammar and functions	17 items	Unit 10, cycle A
Task 4	Grammar and functions	5 items	Unit 11, cycles A + B
Task 5	Grammar and functions	7 items	Unit 12, cycle A
Task 6	Writing	Paragraph	An interesting trip (units 10 and 11)

Table 4 shows that tasks one through five assess grammar and functions, whereas task six assesses writing skills. Tasks one and two focus on the content of unit 9, cycle A, task three focuses on unit 10, cycle A, task four focuses on unit 11, cycles A and B, task five focuses on unit 12, cycle A, and task six focuses on the topic "an interesting trip", related to both units 10 and 11.

As can be observed through table 4, this mid-term test is not uniform regarding its content to be assessed. Five out of six tasks exclusively test grammar and functions, leaving out other components, namely the listening and reading skills, as well as vocabulary. In addition, not all units' cycles are covered: the grammar foci or functions in cycle B of units 9, 10, and 12 are not being assessed at all. The test tasks also contain different number of items. For instance, while tasks two, four, and five contain on average five or seven items each, task three contains seventeen items.

The rubrics in four out of five tasks may also cause misinterpretation on the side of the testee (see appendix A). The rubrics in task one, for instance, "Describe the following people's appearance", could specify what specific features to describe in each picture, as well as the number of sentences required. The rubrics in task five, "Complete the dialogue below" could specify the number of words for each gap. In task six, the composition, whose rubrics read "Write about an interesting trip you have done", does not directly relate to any of the writing topics of units 9 through 12 of the course book, and it lacks extended and clearer directions explaining what specific information is being required. In addition, the number of words to be written is not specified. As in the mid-term test, clarity of information to be expected from the testee is important for the sake of comparing and scoring different students' task performance.

Finally, there is no scoring system written in this mid-term test, and thus it does not allow for any assumptions regarding scoring procedures.

Level 2: final test (New Interchange 1, units 13 through 16): The final test sample selected for the present analysis consists of eight tasks that assess grammar and

functions, and writing (see appendix A). All eight tasks resemble those observed in the first volume of the New Interchange course book. Table 5 shows information regarding the tasks in this final test designed for level 2.

Table 5: Level 2, final test (New Interchange 1, units 13 through 16): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Grammar and functions	5 items	Unit 13, cycle A
Task 2	Grammar and functions	5 items	Unit 13, cycle B
Task 3	Grammar and functions	4 items	Unit 14, cycle A
Task 4	Grammar and functions	4 items	Unit 14, cycle B
Task 5	Grammar and functions	4 items	Unit 15, cycle A
Task 6	Grammar and functions	Not mentioned	Unit 15, cycle B
Task 7	Grammar and functions	4 items	Unit 16, cycle A
Task 8	Writing	Paragraph	Unit 16

According to table 5, tasks one through seven assess grammar and functions: task one focuses on the of unit 13, cycle A, task two focuses on the content of unit 13, cycle B, task three focuses on the content of unit 14, cycle A, task four focuses on the content of unit 14, cycle B, task five focuses on the content of unit 15, cycle A, task six focuses on the content of unit 15, cycle B, and task seven focuses on the content of unit 16, cycle A. The writing task, task eight, focuses the content of unit 15.

Table 2 also shows that in terms of grammar and functions, this final test sample assesses the content of all units it is supposed to assess. However, other components covered in the New Interchange one course book, namely the listening and reading skills, as well as vocabulary, have not been included in the test.

This final test also presents shortcomings regarding the task rubrics (see appendix A). The rubrics of all eight tasks seem vague or lack clearer and extended directions, or even examples, which might hinder the testee's understanding and performance. In task one, for instance, the rubrics are "Write responses for the following statements". It is not specified what kind of information is supposed to be used in the responses. The remaining tasks do also lack additional information or examples regarding what is required from the testee. In addition, specifically in the composition (task eight), rubrics do not specify the number of words to be written.

Unlike its mid-term companion, this test has a score system written on it (see appendix A), and the total score is ten points. However, this test's distribution of points is not uniform. Tasks six, and eight are not worth the same number of points as the other tasks. Task seven (two points) is worth twice as much as tasks one, two, three, four, and five, individually. As a conclusion, it may be observed that the main shortcoming of this final test are the vagueness of its tasks' rubrics, the lack of uniformity in its scoring system, as well as the absence of tasks assessing the following construct components: reading and listening comprehension, and vocabulary.

Level 3: mid-term test (New Interchange 2, units 1 through 4): The mid-term test sample selected for the present analysis consists of five tasks that assess listening comprehension, grammar and functions, and writing (see appendix A) from units 1 through 4 of the New Interchange Two course book. The listening comprehension task (task one) was adapted from the one suggested in test one of the second volume of New Interchange teacher's guide. Two of the four original multiple-choice items were changed and transformed into questions, while the two remaining items, one multiple choice item and the three-step-ordering items remained the same. However, the teacher changed the order of steps. In spite of these changes, the tasks items resemble those observed in listening comprehension tasks in the New Interchange course book. Tasks two, three and four, which were adapted from the teacher's guide test, also resemble those in the course book. The composition's (task three) original main topic, "cooking", was substituted by "evening routine". In addition, the format of task five, a dialog-writing task, has not been observed in the course book. Table 6 presents information regarding the tasks in this mid-term test designed for level 3.

Table 6: Level 3, mid-term test (New Interchange 2, units 1 through 4): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Listening comprehension	9 items	Topics from units 1 through 4
Task 2	Grammar and functions	5 items	Unit 2, cycle B
Task 3	Writing	Paragraph	Unit 4 (topic: evening routine)
Task 4	Grammar and functions	5 items	Unit 3, cycle B
Task 5	Grammar and functions	10 items	Unit 4, cycle A

As can be seen in table 6, task one is a listening comprehension task whose topics refer to those of units 1 through 4. Tasks two, four, and five focus on the grammar and functions content of unit 2 (cycle B), unit 3 (cycle B), and unit 4 (cycle A), respectively. Task three focuses on writing skills whose specifications refer to those of unit 4.

Table 6 reveals that two constructs, namely reading comprehension and vocabulary knowledge, are not assessed in this sample mid-term test. The table also shows that there is not uniformity in the assessment of grammar and functions. There is no task assessing the following content: unit 1 (cycles A + B), unit 2 (cycle A), unit 3 (cycle A), and unit 4 (cycle B).

Examining the test rubrics (see appendix A) reveals that those in task four, "Write a response using *wish* for each statement", could include an example in order to avoid possible misinterpretation from the testee. The rubrics in task five, "Interview a famous person. Ask him/her TEN questions: 5 in the PAST SIMPLE and 5 in the PRESENT PERFECT. Answer all of them", also lack extended directions, such as an explicit topic.

This test contains a score system written on it and thus it has been observed that some tasks are worth more than others (see appendix A). The tasks with the highest scores, for instance, are the listening task (task one, with nine items) and the dialog-writing task (task five), which is worth thirty points. Task two is worth ten points, and tasks three and four are worth fifteen points each. As a result, more weight is put in the assessment of the testee's listening comprehension skill, and the dialog-writing task. We

may thus conclude that the main shortcomings of this mid-term test are the absence of two constructs (namely reading comprehension and vocabulary), and grammar and functions content supposed to be assessed; the rubrics in two tasks, which may lead to misinterpretation by the testee, and the placement of scoring weight in some tasks at the expense of others.

Level 3: final test (New Interchange 2, units 5 through 8): The final test sample selected for the present analysis consists of six tasks that assess listening comprehension, grammar and functions, and writing (see appendix A) from units 5 through 8 of the New Interchange Two course book. As in the mid-term test, in this final test the listening comprehension task (task one) was adapted from the one suggested in test two of the second volume of New Interchange teacher's guide. Two of the four original items were changed from multiple-choice into sentence completion items, while one multiple-choice item and the three-step-ordering items remained the same. In spite of these changes, the tasks items resemble those observed in listening comprehension tasks in the New Interchange Two course book. Task three was literally copied from test two of the New Interchange Two teacher's guide, except for the fact that the teacher did not include the example that accompanies the original task. Tasks three, four, five and six resemble those observed in the course book. Only task two does not resemble any of the tasks in the course book. Table 7 presents information regarding the tasks in this final test designed for level 3.

Table 7: Level 3, final test (New Interchange 2, units 5 through 8): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Listening comprehension	10 items	Topics from units 5 through 8
Task 2	Grammar and functions	Not mentioned	Unit 5, cycle A
Task 3	Grammar and functions	5 items	Unit 6, cycle B
Task 4	Grammar and functions	5 items	Unit 6, cycle A
Task 5	Writing	Paragraph	Unit 5 (topic: letter giving advice to a friend)
Task 6	Grammar and functions	5 items	Unit 7, cycle A

According to table 7, task one is a listening comprehension task whose topics refer to those of units 5 through 8. Tasks two, three, four, and six focus on the grammar and functions content of unit 5 (cycle A), unit 6 (cycles A and B), and unit 7 (cycle A), respectively. Task five focuses on writing skills whose specifications refer to those of unit 5.

Table 7 reveals that, similarly to the mid-term test, two constructs in this sample final test, namely reading comprehension and vocabulary knowledge, are also not assessed. The table also shows that there is not uniformity in the assessment of grammar and functions. There is no task assessing the contents of unit 7 (cycle B) and unit 8 (cycles A and B).

A detailed examination of the test tasks (see appendix A) reveals that although the task topics remain faithful to those in the course book, the rubrics in tasks two, four, five, and six lack the additional information and cues and examples that usually accompany the rubrics in the course book tasks. Task four (gap-filling), for instance, whose rubrics read "Use the right preposition to complete the sentences below", does not provide a list of the words to be used in the gaps, whereas task six could also provide additional cues (such as useful vocabulary) in order to ensure the testee's optimal performance of the task.

This test also contains a score system written on it and thus it has been observed that some tasks are worth more than others (see appendix A). The tasks with the highest scores are the listening task (task one, worth twenty-five points), task two (worth twenty points), and the writing task (task five, also worth twenty points). Tasks three is worth fifteen points, and tasks four and six are worth ten points each. As in the mid-term test, in this final test sample more weight is put in the assessment of the testee's listening comprehension skill, the dialog-writing task, and the writing task. In conclusion, the main shortcomings of this final test are the absence of two constructs (namely reading comprehension and vocabulary), the absence of part of the grammar and functions content supposed to be assessed, the vagueness of rubrics in four out of six tasks that

may lead to misinterpretation by the testee, and also the placement of scoring weight in some tasks at the expense of others.

Level 4: mid-term test (New Interchange 2, units 9 through 12): The mid-term test sample selected for the present analysis consists of six tasks that assess listening comprehension, grammar and functions, and writing (see appendix A) from units 9 through 12 of the New Interchange Two course book. The listening comprehension task (task one) was adapted from the one suggested in test three of the second volume of New Interchange teacher's guide. Besides the three original true/false items, eight comprehension questions were added. However, in spite of these changes, the tasks items resemble those observed in listening comprehension tasks in the New Interchange Two course book. Task two, the writing task, was adapted from the writing task of test three in the New Interchange Two teacher's guide. Tasks one, two, four and five resemble those observed in the course book, whereas tasks three and six do not. Table 8 presents information regarding the tasks in this mid-term designed for level 4.

Table 8: Level 4, mid-term test (New Interchange 2, units 9 through 12): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Listening comprehension	12 items	Topics from units 9 through 12
Task 2	Writing	Paragraph	Unit 9 (topic: your past, present, and future)
Task 3	Grammar and functions	7 items	Unit 10, cycle A
Task 4	Grammar and functions	5 items	Unit 9, cycle B
Task 5	Grammar and functions	5 items	Unit 12, cycle A
Task 6	Grammar and functions	Not mentioned	Unit 12, cycle B

As can be observed from table 8, task one is a listening comprehension task whose topics refer to those of units 9 through 12. Tasks three, four, five, and six focus on the grammar and functions content of unit 10 (cycle A), unit 9 (cycle B), unit 12 (cycle A), and unit 12 (cycle B) respectively. Task two focuses on writing skills whose specifications refer to those of unit 9.

Table 8 shows that two constructs in this sample mid-term test, namely reading comprehension and vocabulary knowledge, are not assessed. In addition, there is not uniformity in the assessment of grammar and functions. There is no task assessing the contents of unit 10 (cycle B), and unit 11 (cycles A and B).

Examining the test tasks (see appendix A) reveals that the rubrics in tasks three, four, and five, lack the information or examples, which may lead to misinterpretation by the testee. Task four, for instance, whose rubrics read "Complete these sentences with your own information", and task five, whose rubrics read "Complete these sentences", do not provide any additional information regarding what exactly the testee is supposed to complete the sentences with. In addition, an example in each of these tasks would ensure the testee's understanding of what is required.

By examining the score system written on this mid-term test it has been observed that some tasks are worth more than others (see appendix A). The tasks with the highest scores are the listening task (task one, worth thirty-two points) and task two (worth nineteen points). Tasks three is worth fourteen points, task four is worth ten points, task five is worth twelve points, and task six is worth fifteen points. The result is that in this final test sample more weight is put in the assessment of the testee's listening comprehension and the writing skill. We may thus conclude that the main shortcomings of this final test are the absence of two constructs (namely reading comprehension and vocabulary), the fact that part of the grammar and functions content is not assessed, and the rubrics in four out of six tasks that may lead to misinterpretation by the testee, and finally the placement of scoring weight in some tasks at the expense of others.

Level 4: final test (New Interchange 2, units 13 through 16): The final test sample selected for the present analysis consists of six tasks that assess listening comprehension, grammar and functions, and writing (see appendix A) from units 13 through 16 of the New Interchange Two course book. In this final test the listening

comprehension task (task one) was also adapted from the one suggested in test four of New Interchange Two teacher's guide. Besides the six original multiple choice items, four comprehension questions and four sentence completion items were added, but in spite of these changes, the tasks items resemble those observed in listening comprehension tasks in the New Interchange Two course book. Tasks three and five have been taken from the teacher's guide of New Interchange Two, and tasks two, four, and six resemble those observed in the course book. Table 9 presents information regarding the tasks in this final test designed for level 4.

Table 9: Level 4, final test (New Interchange 2, units 13 through 16): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Listening comprehension	14 items	Topics from units 13 through 16
Task 2	Writing	Paragraph	Units 13 and 15
Task 3	Grammar and functions	7 items	Unit 13, cycle A
Task 4	Grammar and functions	6 items	Unit 14, cycle B
Task 5	Grammar and functions	5 items	Unit 15, cycle A
Task 6	Grammar and functions	2 items	Unit 15, cycle B

According to table 9, task one is a listening comprehension task whose topics refer to those of units 13 through 16. Tasks three, four, five, and six focus on the grammar and functions content of unit 13 (cycle A), unit 14 (cycles B), and unit 15 (cycle A), and unit 16 (cycle B) respectively. Task two, which focuses on writing skills, allows the testee to choose one of the three valid topics suggested (see appendix A). Topic "a" refers to specifications of unit 15, topics "b" and "c" refer to specifications of units 13. Topic "c", however, constitutes another alternative to option "b", and requires the testee to write about a book (or reader) previously read in class.

As it may be observed from table 9, in this final test two constructs in this sample final test, namely reading comprehension and vocabulary knowledge, are not assessed. The table also shows the lack of uniformity in the assessment of grammar and functions. There is no task assessing the contents of unit 13 (cycle B), unit 14 (cycle A, and unit 16 (cycles A and B).

A detailed examination of the test tasks (see appendix A) reveals the rubrics in all tasks seem complete, although the inclusion of examples could contribute to the testee's understanding of what is required. Task two, the writing, whose rubrics are "Choose one of the topics below, and write no less than 100 words", include the number of words required, but do not provide any extended instructions regarding the organization of the compositions.

This test also contains a score system written on it and thus it has been observed that some tasks are worth more than others (see appendix A). The tasks with the highest scores are the listening task (task one, worth thirty-three points) and task two (worth twenty-five points). Tasks three is worth seven points, task four is worth twelve points, task five is worth fifteen points, and task six is worth eight points. In conclusion, the main shortcomings of this final test are the absence of two constructs (namely reading comprehension and vocabulary), the absence of part of the grammar and functions content supposed to be assessed, and also the placement of scoring weight in some tasks at the expense of others.

Level 5: mid-term test (New Interchange 3, units 1 through 4): This mid-term test sample selected for the present analysis is divided into two parts: listening and writing (see appendix A), which consist of nine tasks, identified by letters, assessing listening comprehension, vocabulary, grammar and functions, and writing. The listening comprehension task (task A, part I), the vocabulary task (Task A, part II), and a grammar and functions task (tasks C, part II) were adapted from similar tasks in test one in the teacher's guide of New Interchange Three. In the listening comprehension task (task A, part I), for instance, of the four original multiple-choice items, three alternative ones were created while another was substituted by two comprehension questions, but despite these changes, the items in this listening comprehension task have been observed in the New Interchange Three course book. Tasks B, C, D, and E, however, do not resemble any of the tasks observed in the New Interchange Three course book.

Table 10 shows information regarding each task in this mid-term sample test designed for level 5.

Table 10: Level 5, mid-term test (New Interchange 3, units 1 through 4): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task A (part I)	Listening comprehension	5 items	Topics from units 1 through 4
Task A (part II)	Vocabulary	5 items	Units 1 and 2
Task B (part II)	Grammar and functions	Not mentioned	Unit 1, cycle B
Task C (part II)	Grammar and functions	5 items	Unit 2, cycle B
Task D (part II)	Vocabulary	5 items	Unit 3
Task E (part II)	Grammar and functions	11 items	Unit 4, cycle A; unit 3, cycle A; and unit 2, cycle A
Task F (part II)	Grammar and functions	12 items	Unit 4, cycle A
Task G (part II)	Grammar and functions	5 items	Unit 4, cycle B
Task H (part II)	Writing	Paragraph	Unit 2, cycle A; and unit 4, cycles A + B

According to table 10, task A (part I) focuses on the listening comprehension, whose main topics relate to units 1 through 4 of the course book. Task A (part II) focuses on the vocabulary content of units 1 and 2 of the course book, task D focuses on the vocabulary content of unit 3 of the course book. Tasks B, C, E, F, and G focus on the grammar and functions content of units 1 through 4 of the course book, and task H focuses on the writing skill with two optional topics from the course book: on that refers to specifications in unit 4, cycles A and B, and another that refers to specifications in unit 3, cycle A.

Table 10 also reveals that the grammar and function, and vocabulary content (part II) of this test is not assessed with uniformity. Task E, which contains eleven items, for instance, aims to assess the grammar and functions contents of units 4 (cycle A) in seven items, unit 3 (cycle a) in three items, and unit 2 (cycle A) in one item. Task F assesses the content of Unit 4, cycle A in twelve items. The following grammar and functions content is not assessed: unit 1 (cycle A), unit 3 (cycle B), as well as the vocabulary of unit 4. In addition, this mid-term test does not assess reading comprehension.

Examining the test reveals a few other problems regarding the task formats (see appendix A). The two tasks assessing vocabulary knowledge may affect negatively in the testee's performance. Task A (part II), for instance, was adapted from a similar task in test one of the New Interchange Three teacher's guide and assesses vocabulary of units one and two of the course book. In its original version, however, this is a multiple-choice task, whereas in this test this was transformed into a gap-filling task where the original possible options were not included allowing the possibility to use words other than those intended.

In tasks C, D, and E (see appendix A) the rubrics lack extended directions, which may lead to the testee's misinterpretation of the task. In task C, for example, whose rubrics are "Write sentences that have the same meaning", it is not clear what exactly this task requires from the testee. It may thus be concluded that this mid-term test is limited in terms of the grammar and functions, and vocabulary content it is supposed to assess, the task rubrics and formats may hinder the testee's performance, and it does not assess reading comprehension.

This test does not present a scoring system, which makes it difficult to assume how much each item or task is actually worth.

Level 5: final test (New Interchange 3, units 5 through 8): This final test sample selected for the present analysis is also divided into two parts: listening and writing (see appendix A), which consist of nine tasks, identified by letters, assessing listening comprehension, vocabulary, grammar and functions, reading comprehension, and writing. The listening comprehension task (task A, part I), the vocabulary task (Task A, part II), and five grammar and functions tasks were adapted (tasks B, C, D, and F, part II) or cut and pasted (task G, part II) from tasks in test two in the teacher's guide of New Interchange Three. In the listening comprehension task (Task A, part I), for instance, three of the four original multiple-choice items were substituted by three sentence completion items, but despite these changes, the items in this listening comprehension

task have been observed in the New Interchange Three course book. One task out of eleven (tasks E), however, does not resemble any of the tasks observed in the New Interchange Three course book. Table 11 shows information regarding each task in this mid-term sample test designed for level 5.

Table 11: Level 5, final test (New Interchange 3, units 5 through 8): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task A (part I)	Listening comprehension	4 items	Topics from units 5 through 8
Task A (part II)	Vocabulary	6 items	Unit 5
Task B (part II)	Grammar and functions	3 items	Unit 5, cycle B
Task C (part II)	Grammar and functions	4 items	Unit 6, cycles A + B
Task D (part II)	Grammar and functions	6 items	Unit 6, cycle A
Task E (part II)	Grammar and functions	11 items	Unit 7, cycles A + B
Task F (part II)	Grammar and functions	5 items	Unit 8, cycle A
Task G (part II)	Reading comprehension	4 items	Unit 8 (topic: "developing good habits")
Task H (part II)	Writing	Paragraph	Unit 6, cycles A + B and unit 4, cycle A

According to table 11, task A (part I) focuses on the listening comprehension, whose main topics relate to units 5 through 9 of the course book. Tasks A (part II) focuses on the vocabulary content of unit 5 of the course book, tasks B, C, D, E, and F focus on the grammar and functions content of units 5 through 9 of the course book, task G focuses on reading comprehension, whose topic is "developing good reading habits", and task H focuses on the writing skill with two optional topics from the course book: on that refers to specifications in unit 6, cycles A and B, and another that refers to specifications in unit 5, cycle B.

Table 11 also shows that the grammar and function, and vocabulary content (part II) of this test is not assessed with uniformity. Task E, which contains eleven items, for instance, aims to assess the grammar and functions content of unit 7 (cycles A and B). The content of cycle A, however, is assessed in two items, while the content of cycle B is assessed in nine items. The vocabulary task (task A, part II) assesses the content of unit 5 only. The following grammar and functions content is not assessed: unit 5 (cycle A), unit 8 (cycle B), as well as the vocabulary of units 6, 7, and 8.

Examining the test reveals few other problems regarding task rubrics. In tasks C, D, E, and F (see appendix A) the rubrics lack extended directions, which might confuse the testee. In task D, for example, whose rubrics are "Rewrite the sentences in a different way", it is not clear how the sentences should be rewritten. An example sentence in this task could avoid misinterpretation problems. It may thus be concluded that this mid-term test presents problems in terms of the grammar and functions, and vocabulary content it is supposed to assess, as well as in terms of incomplete task rubrics that may hinder the testee's performance.

As in the mid-term test for level 5, this final test does not present a scoring system, which makes it difficult to assume how much each item or task is actually worth.

Level 6: mid-term test (New Interchange 3, units 9 through 12): The mid-term test sample selected for the analysis in the present study is composed of ten tasks assessing listening and reading comprehension, and grammar and functions (see appendix A). Five out of ten tasks (tasks one, three, four, seven, and ten) have been taken from test three in the New Interchange Three teacher's guide, with some adaptations. In task 1, for instance, the multiple-choice listening task, the teacher changed the order of options. In task 4, the teacher changed the original multiple-choice items into gap-filling items. The remaining tasks (tasks 2, 5, 6, 8, and 9) resemble tasks observed in the New Interchange Three course book. Table 12 presents information regarding each task in the mid-term test sample designed for level 6.

Table 12: Level 6, mid-term test (New Interchange 3, units 9 through 12): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Listening comprehension	4 items	Units 9 through 12
Task 2	Grammar and functions	3 items	Unit 9, cycle A
Task 3	Grammar and functions	5 items	Unit 9, cycle B
Task 4	Grammar and functions	4 items	Unit 10, cycle A
Task 5	Grammar and functions	6 items	Unit 10, cycle B
Task 6	Grammar and functions	4 items	Unit 11, cycle B
Task 7	Grammar and functions	6 items	Unit 11, cycle A; unit 12, cycle A
Task 8	Grammar and functions	4 items	Unit 11, cycle B
Task 9	Grammar and functions	4 items	Unit 12, cycle B
Task 10	Reading comprehension	6 items	Unit 12 (topic: "reading is fun")

Table 12 shows that tasks one focuses on listening comprehension, whose topics refer to those of units 9 through 12 of the course book; tasks two through nine focus on the grammar and functions content of units 9 through 12 of the course book, respectively, and task ten assesses reading comprehension, and its main topic is "reading is fun", related to unit 12 of the course book.

Table 12 also reveals that, although all grammar and functions content is assessed in this mid-term test, it does not include the assessment of two constructs, namely vocabulary and writing.

In terms of test rubrics, however, this mid-term test presents a few shortcomings (see appendix A), such as those in tasks five and seven. In task five, for instance, whose rubrics read "Complete with the correct verb tense", it could be specified what verb tenses the testee will use, in both the active and passive voice. In task seven, whose rubrics are "Complete the sentences with your own information", the teacher could have provided an example in order to ensure the testee understands what is required. In sum, the main drawbacks of this mid-term test are the absence of two constructs (vocabulary and writing), and the vague rubrics in two of the ten tasks. Regarding scoring, the absence of a scoring or marking system written on the test hinders any possible assumptions about score distribution among the test tasks.

Level 6: final test (New Interchange 3, units 13 through 16): The final test sample selected for the analysis in the present study is composed of nine tasks assessing listening and reading comprehension, grammar and functions, and vocabulary (see appendix A). All nine tasks have been taken from test four in the New Interchange Three teacher's guide, with some adaptations. In task 1, 2, and 3 for instance, the teacher either changed the order of the multiple-choice options, or the items. Nevertheless, all tasks resemble those observed in the New Interchange Three course book. Table 13 presents information regarding each task in the final test sample designed for level 6.

Table 13: Level 6, final test (New Interchange 3, units 13 through 16): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Listening comprehension	4 items	Units 13 through 16
Task 2	Grammar and functions	4 items	Unit 13, cycle A
Task 3	Grammar and functions	4 items	Unit 13, cycle B
Task 4	Grammar and functions	4 items	Unit 15, cycle B
Task 5	Grammar and functions	4 items	Unit 14, cycle B
Task 6	Grammar and functions	4 items	Unit 15, cycle A
Task 7	Vocabulary	4 items	Unit 16
Task 8	Grammar and functions	5 items	Unit 16, cycle B
Task 9	Reading Comprehension	6 items	Unit 14 (topic: the movie camera)

As can be observed from Table 13, tasks one focuses on listening comprehension, whose topics refer to those of units 13 through 12 of the course book; tasks two, three, four, five, six, and eight focus the grammar and functions content of units 9 through 12 of the course book, respectively; task seven focuses on the assessment of vocabulary of unit 16, and task nine assesses reading comprehension, and its main topic is 'reading is fun", related to unit 14 of the course book.

Table 12 also reveals that there is not uniformity in terms of the content to be assessed. For instance, not all grammar and functions content is assessed in this mid-term test, such as that of unit 14 (cycle A), and unit 16 (cycle A). The vocabulary task does not assess the vocabulary content of units 13, 14, and 15. In addition, this final test does not include the assessment of writing skills.

In terms of test rubrics, however, eight out of nine tasks provide detailed instructions that do not seem to cause misinterpretation by the testee (see appendix A). Only in task eight, however, whose rubrics are "Complete these sentences with your own information", the teacher could have included extended directions. Otherwise, in some items (see item five, for instance) the testee might still complete the sentences correctly with structures other than those that are really assessed. We may thus conclude that the main drawbacks of this final test are the absence of one construct (writing skills), and the absence of some grammar and functions, and vocabulary content supposed to be assessed.

Regarding scoring, as in the mid-term test for level 6, the absence of a scoring or marking system written on the test hinders any possible assumptions about score distribution among the test tasks.

Level 7: mid-term test (Passages one, units 1 through 3): The mid-term test sample selected for the analysis in the present study is composed of ten tasks, identified by letters, assessing grammar and functions, reading comprehension and writing (see appendix A). Six out of seven tasks resemble those observed in the Passages One course book. Table 14 presents information regarding each task in the mid-term test sample designed for level 7.

Table 14: Level 7, mid-term test (Passages 1, units 1 through 3): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task A	Grammar and functions	5 items	Unit 1, lesson A
Task B	Grammar and functions	4 items	Unit 1, lesson B
Task C	Grammar and functions	4 items	Unit 2, lesson A
Task D	Grammar and functions	5 items	Unit 2, lesson B
Task E	Grammar and functions	4 items	Unit 3, lesson A
Task F	Grammar and functions	4 items	Unit 3, lesson B
Task G	Reading and Writing	Paragraphs	Unit 2 (topic: "language learning")

According to table 14, tasks A through F focus the grammar and functions content of units 1 through 3 of the course book, and task G is a tasks that integrates the

assessment of both reading comprehension and writing skills, and its main topic is "language learning", related to unit 2 of the cou rse book. Table 14 also shows that although all grammar and functions content of the three units is assessed, this mid-term test does not assess two constructs, namely vocabulary and listening comprehension.

This mid-term test sample presents shortcomings with respect to its task rubrics (see appendix A). In A, for instance, whose rubrics are "Complete the sentences below with information of your own", the testee might not understand what exactly is supposed to be done. In all items it is possible to complete the sentences with nouns, instead of *verbs in the infinitive or gerund*, the main focus of task A. In tasks B and D, for instance, the absence of examples does not make clear how exactly the items are supposed to be completed.

Problems regarding the assessment of grammar and functions content and the writing skills have also been observed (see appendix A). Task E, for instance, assesses the use of non-defining relative clauses only, while the main purpose of the course book unit and lesson it refers to (Unit 3, lesson A) is to teach the contrast of both *defining* and *non-defining relative clauses*. Task G, a task integrating the assessment of the reading and writing skills, requires the student to write two short forty-word paragraphs in reaction to the reading passage. Although the main topic is the same as that of unit 2 ("learning a language"), this task does not fully explore the central writing process of unit 2: the writing of topic sentences. We may thereby conclude that the main drawbacks of this mid-term test are the absence of two constructs (listening comprehension and vocabulary), the lack of extended directions in the rubrics of two of the seven tasks, and the assessment of grammar and functions content and writing skills in two of the seven tasks.

Again, the absence of a scoring or marking system written on this test hinders any possible assumptions about score distribution among the test tasks.

Level 7: final test (Passages one, units 4 through 6): The final test sample selected for the analysis in the present study is composed of six tasks, identified by letters, assessing grammar and functions, vocabulary, and writing (see appendix A). Only two out of six tasks resemble those observed in the Passages One course book. Table 15 presents information regarding each task in the mid-term test sample designed for level 7.

Table 15: Level 7, final test (Passages 1, units 4 through 6): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task A	Grammar and functions	12 items	Unit 4, lessons A + B
Task B	Grammar and functions	4 items	Unit 6, lesson A
Task C	Grammar and functions	9 items	Unit 6, lesson B
Task D	Grammar and functions	15 items	Unit 5, lesson B
Task E	Vocabulary	20 items	Objects
Task F	Writing	Paragraph	English in the professional or private life

According to table 15, tasks A through D focus the grammar and functions content of units 4 through 6 of the course book, task E focuses on the assessment of vocabulary (sorted out objects), and task F focuses on the assessment of the writing skills, and its main topic is "English in the professional and private life".

As can be observed through table 15, this final test does not assess the following constructs: listening comprehension and reading comprehension. In terms of content, although there is a vocabulary task (task E), this final test does not assess any of the lexical areas presented in units 4 through 6. In addition, it does not assess the grammar and functions content of unit 5, lesson A.

This final test sample also presents shortcomings with respect to its task rubrics (see appendix A). In tasks A, B, and D the rubrics lack extended directions, which might lead to misinterpretation from the testee. In task B, for instance, the rubrics ("Think of how Brazilians behave in typical social occasions and then complete the sentences below") do not make clear what exactly is required from the testee. The rubrics in task D, "Fill in the blanks to make conditional sentences", do not specify what kind of *conditional sentence* to be used. In tasks A, B, C and D, for instance, the

absence of examples does not make clear how exactly the task items are supposed to be completed.

Other shortcomings refer to the task formats per se (see appendix A). Task E assesses vocabulary by requiring the translation of twenty isolated words from English to Portuguese. This type of task has not been observed in the Passages course book and the lexical areas to which the words belong to, are not part of the vocabulary content to be assessed in units 4 through 6. In addition, task F, the composition, does not follow the topic or the writing process practiced in any of the units that cover this test (units 4 through 6). While in the course book students are supposed to have practiced the writing of topic sentences and paragraphs, the composition topic in task F reveals to be out of the intended context. In sum, the main drawbacks of this final test are the absence of two constructs (listening and reading comprehension), the lack of extended directions in the rubrics of three out of six tasks, and the absence of the assessment of grammar and functions content of unit 5, cycle A.

With respect to scoring, the absence of a scoring or marking system written on this test hinders any possible assumptions about score distribution among the test tasks.

Level 8: mid-term test (Passages one, units 7 through 9): The mid-term test sample selected for the analysis in the present study is composed of six tasks assessing grammar and functions, vocabulary, and writing skills (see appendix A). Five out of six tasks resemble those observed in the Passages One course book. Table 16 presents information regarding each task in the mid-term test sample designed for level 8.

Table 16: Level 8, mid-term test (Passages 1, units 7 through 9): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Grammar and functions	4 items	Unit 7, lesson A
Task 2	Vocabulary	5 items	Units 7, 8, and 9
Task 3	Grammar and functions	8 items	Unit 7, lesson B
Task 4	Grammar and functions	5 items	Unit 8, lesson B
Task 5	Grammar and functions	5 items	Unit 8, lesson A; unit 9, lesson A
Task 6	Writing	Paragraph	Unit 8 (topic: a significant event in your past)

According to table 16, tasks one, three, four, and five focus on the grammar and functions content of units 7 through 9 of the course book. Task 2 focuses on the vocabulary of units 7 through 9, and task 6 focuses on the assessment of writing skills practiced in Unit 8. Table 16 also reveals that the grammar and functions content of unit 9, lesson B is not assessed. In addition, this mid-term test does not assess two constructs, namely reading and listening comprehension.

Although the tasks in this mid-term test require the testee to produce more language (that is, they assess the testee's knowledge of the test content by requiring subjective answers), they present shortcomings regarding the rubrics and the task format in two out of five tasks (see appendix A). Task two, a vocabulary task in which the test taker is supposed to explain, give examples, or define specific words, has not been observed in the course book. Its rubrics, which read "choose FIVE of the instructions or questions containing words or expressions in **bold**. Don't use more than 40 words in each answer", fails to provide extended details on how the testee's knowledge of the words or expressions is going to be measured. In task five, which requires the testee to answer questions with his or her own words, the rubrics do not guarantee the testee will make use of the grammatical and functional structures that the five open questions elicit (question "a", for instance, elicits the use of *reduced relative clauses*, the grammatical content of unit 8, lesson A). Moreover, although this task seems to allow a high degree of interaction between the testee and the task itself, main doubt remains on how the answers would be scored.

Attention should also be drawn on task six (see appendix A). This task is literally a multi-paragraph composition presented in unit 8, lesson A. However, the drawback of the way it is presented in this test seems to lie in the fact that it is performed and revised before they sit this test, which in this case might not be characterized as an assessment task. In other words, students do not perform the composition task in the test situation, on the contrary, the teacher uses the final draft -

produced by students in class or at home before the test - as the instrument for the assessment of the writing skills.

It is thus possible to conclude that, despite assessing language through tasks which elicit subjective responses from the testee, the main drawbacks of this mid-term test are the absence of two constructs (listening and reading comprehension, the lack of extended directions in the rubrics of two out of six tasks, and the absence of tasks assessing the grammar and functions content of unit 9, lesson B.

This test contains a scoring system written on it, but although it evenly assesses the vocabulary content of all three units, a few problems of weight have been observed in the grammar and functions tasks: more weight is placed in tasks two (vocabulary task) and five (short questionnaire), as opposed to task one, which is the least weighted of all six tasks. In addition, the number of items in each task varies as well: task one, for instance, has four items, whereas task three has eight items.

Level 8: final test (Passages one, units 10 through 12): The final test sample selected for the analysis in the present study is composed of eight tasks assessing grammar and functions, vocabulary, and writing skills (see appendix A). Six out of eight tasks resemble those observed in the Passages One course book. Table 17 presents information regarding each task in the final test sample designed for level 8.

Table 17: Level 8, final test (Passages 1, units 10 through 12): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Grammar and functions	3 items	Unit 10, lesson A; unit 12, lesson B
Task 2	Grammar and functions	2 items	Unit 11, lesson A
Task 3	Grammar and functions	2 items	Unit 11, lesson A
Task 4	Grammar and functions	2 items	Unit 10, lesson B
Task 5	Writing	Paragraph	Unit 10, 11, and 12 (optional topics)
Task 6	Grammar and functions	2 items	Unit 12, lesson A
Task 7	Grammar and functions	2 items	Unit 11, lesson B
Task 8	Vocabulary	3 items	Units 10 through 12

According to table 17, tasks one, two, three, four, six, and seven focus on the grammar and functions content of units 10 through 12 of the course book. Task 5, which

focuses on the writing skills, deals with topics presented in units 10 through12, and task eight focuses on the vocabulary of units 10 through 12. Table 17 also reveals that this final test does not assess two constructs, namely reading and listening comprehension.

Although the tasks in this final test also require the testee to produce more language, rather than eliciting knowledge through multiple-choice or gap-filling type of tasks, it presents shortcomings regarding the rubrics and the task format in three out of eight tasks (see appendix A). Task one, whose rubrics are "Answer the following questions with personal information", could provide an example in order to avoid misinterpretation by the testee. In addition, in items "a" and "b", which require an answer with *noun clauses containing relative clauses,* the testee might provide an answer with *nouns* instead of *clauses*. The type of questions in the items of task five (short-paragraph composition) have been observed in *oral production* activities of the course book, not in *written* ones. In addition, none of the items assess the original writing process presented in each of the course book units 10 through 12 that this test is supposed to cover. Task eight, a vocabulary task in which the test taker is supposed to explain or define specific words or expressions, has not been observed in the course book. Its rubrics, which read "choose THREE of the instructions or questions containing words or expressions in **bold**", as in the mid-term test, fail to provide extended details on how the testee's knowledge of the words or expressions is going to be measured. We may thereby conclude that the main drawbacks of this mid-term test are the absence of two constructs (listening and reading comprehension), and the lack of extended directions in the rubrics of two out of eight tasks. Moreover, although tasks such as one, three, and six, for instance, seem to allow a high degree of interaction between the testee and the task itself, one main doubt remains on how the answers to their items would be scored.

The scoring system of this test sample is written on the rough copy that was used for the present analysis. Despite the uniformity in terms of numbers of items for each task, as with the mid-term test, some tasks are worth more than others. While more

emphasis is placed in tasks one (short questionnaire, worth twenty points), task five (one-paragraph composition, worth fifteen points), and task eight (vocabulary knowledge task, worth fifteen points), the remaining tasks are worth ten points each.

Level Adv. 1: mid-term test (Passages two, units 1 through 3): The mid-term test sample selected for the analysis in the present study is composed of five tasks assessing grammar and functions, vocabulary, listening and reading comprehension, and writing skills (see appendix A). Task 4, in particular, assesses both reading and listening skills in an integrated manner. It consists of an attached photocopy of an article (which has also been audio recorded) taken from a magazine whose source is unknown. The teacher was immediately enquired about the source of this material – both the photocopy and the audio recorded material. However, the teacher claimed she did not remember where it had been taken from. However, she reported that the listening task consisted of a fac-simile copy of the article in which some words were randomly erased using correction fluid. When she was told that this latter version with erased words was needed for the present analysis, she also claimed that despite her efforts she was not able to find it. As a result, for task four, only the reading comprehension questions and the photocopy of the article (without the gaps) were made available (see appendix A).

Regarding the extent to which the tasks in this final test resemble those in the Passages Two course book, only task one shows resemblance to task formats observed in the Passages Two course book. Table 18 presents information regarding each task in the mid-term test sample designed for level Adv. 1.

Table 18: Level Adv. 1, mid-term test (Passages 2, units 1 through 3): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Grammar and functions	5 items	Unit 2, lesson B, unit 3, lesson B
Task 2	Grammar and functions	5 items	Unit 1, lesson A
Task 3	Vocabulary	5 items	Unit 1
Task 4	Reading and Listening comprehension	Reading: 5 items; listening: unknown.	Unit 2 (topic: "clothes and appearance")
Task 5	Writing	Paragraph	Units 1 through 3 (optional topics)

As can be observed from table 18, tasks one and two focus on the grammar and functions content of units 1 through 3 of the Passages course book, task 3 focuses on the vocabulary of unit 1, task 4 focuses on reading and listening comprehension whose topic is related to unit 2, and task 5 focuses on the assessment of the writing skills with topics covered in units 1 through 3 of the passages two course book. Table 18 also reveals that the grammar and functions content of unit 1 (lesson B), unit 2 (lesson A), and unit 3 (lesson A), as well as the vocabulary content of units 2 and three, are not assessed.

This mid-term test presents shortcomings regarding the rubrics and the task format in all tasks (see appendix A). Task one, for instance, focuses on the use of *cleft sentences with "what"*, as well as *superlative adjectives*, and its rubrics are "Complete these sentences using personal experiences". Although it allows the test taker to express his or her opinion, it lacks extended directions on what exactly is required from the testee. In task three, which requires the testee to explain the meaning of vocabulary related to unit 1 (*adjectives describing incidents or* events), the teacher could have included an example in order to avoid misinterpretation from the testee. In addition, it fails to provide extended details on how the testee's knowledge of the words or expressions is going to be measured.

The integrated task four, assessing listening and reading comprehension, seems to be the one that stands out in terms of shortcomings (see appendix A). The lack of the appropriate listening material used in the task did not allow me to analyze the task. However, if erasing words from the text made this gap-filling task, at least it is safe to assume that gap-filling listening tasks are not common in the Passages Two course book. Still, one may wonder what criterion was adopted for choosing the text and selecting the words to be erased. As for the follow up reading comprehension questions, they differ from those occurring in the course book reading activities in that they only check facts. The reading comprehension tasks in the course book, as the level is

"advanced", either consist of true/false statements, or questions stimulating thinking and opinion giving.

The writing task (task five), or composition, presents the test taker with three options (see appendix A). Although all three topics are valid, that is, they are the same as those in units 1 through 3, of the Passages Two course book, "a" and "c" have actually been practiced as oral activities (discussions), not as writing tasks. The only exception is option "b", which is closest to the writing process and topic suggested in unit two, namely "writing a composition about a personal belief; giving examples to support a thesis statement". As conclusion, it is safe to observe that this tests presents the following shortcomings: the formats of five out of six tasks have not been observed in the Passages Two course book, and task rubrics of tasks one, two, three, for instance, might lead to misinterpretation by the testee. In addition, not all the grammar and functions, and vocabulary content is assessed

The absence of a scoring system in this test does not allow for any assumptions regarding the score distribution among test tasks.

Level Adv. 1: final test (Passages two, units 4 through 6): The final test sample selected for the analysis in the present study is composed of four tasks assessing grammar and functions and writing skills (see appendix A). Only tasks two and four resemble task formats observed in the Passages Two course book. Table 19 presents information regarding each task in the final test sample designed for level Adv. 1.

Table 19: Level Adv. 1, final test (Passages 2, units 4 through 6): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Presumably writing	5 items	Topics: Units 4 and 5
Task 2	Grammar and functions	5 items	Unit 5, lesson B
Task 3	Grammar and functions	5 items	Unit 6, lesson A
Task 4	Writing	Paragraph	Unit 6 (optional topics)

As can be observed from table 19, tasks one, presumably assesses writing skills, as none of its task items focus on a specific grammar point or function, or even

vocabulary (see appendix A). In this task only the topics of the items (the questions) are related to the units this test is supposed to assess (units 4 through 6). Tasks two and three focus on the grammar and functions content of units 5 and 6 of the Passages course book, and task 4 focuses on the assessment of the writing skills with topics covered in unit 6. Table 19 also reveals that the grammar and functions content of unit 4 (lessons A and B) is not assessed. In addition the following constructs are absent: reading and listening comprehension, and vocabulary.

This final test presents shortcomings regarding the rubrics and the task format in three tasks (see appendix A). In task one, for instance, whose rubrics are "Answer the following questions using your personal view to the points explored", it is not possible to determine what construct or content is assessed. It consists of five general questions whose topics are the same as those encountered in units 4 and 5 of the Passages Two course book, but items do not assess any specific grammar focus or functions related to the units that this test is supposed to cover (units 4 through 6). The task rubrics and format leads us to presume that it assesses writing skills.

Task two, whose rubrics are "Use Negative Adverbs like *never, hardly ever, rarely, seldom* to rewrite the following sentences (see what changes are necessary)", resembles one task observed in the Passages Two course book, but a few changes have constrained its original purpose: the original task in the book requires the test taker to change the position of a negative adverb from the middle to the beginning of a given sentence (ex: *Quiz shows* ***seldom*** *require participants to know a subject in depth*, which should then be transformed into ***Seldom*** *do quiz shows require participants to know a subject in depth*). However, in the test task, the adverbs, which in the original task appear in bold, were not included in the cue sentences. As a result the task taker's performance might be extremely hindered by these changes. In addition, the rubrics do not specify whether the adverbs should be used in the beginning or middle of sentences.

In task three, not only is its format unfamiliar to the test taker, since it does not resemble any tasks in the course book, but also the lack of extended directions and an example makes one wonder what this task is actually assessing (see appendix A).

In the composition (task four), although the three topics are dealt in unit 6, none of them explores any of the writing processes presented in the three units (such as *restating a thesis, writing a book report*, and *writing a classification essay*). As conclusion, it is clear that that these tests present shortcomings in terms of task rubrics and formats, which might lead to misinterpretation by the testee. In addition, apart from the fact that this final test does not assess reading and listening comprehension, not all the grammar and functions, and vocabulary content is assessed.

As in the mid-term test for this level (Adv. 1), the absence of a scoring system in this test does not allow for any assumptions regarding the score distribution among test tasks.

Level Adv. 2: mid-term test (Passages two, units 7 through 9): The mid-term test sample selected for the analysis in the present study is composed of six tasks assessing grammar and functions, listening comprehension, and writing skills (see appendix A). Only tasks four and five resemble task formats observed in the Passages Two course book. Table 20 presents information regarding each task in the mid-term test sample designed for level Adv. 2.

Table 20: Level Adv. 2, mid-term test (Passages 2, units 7 through 9): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Grammar and functions	5 items	Unit 7, lesson A
Task 2	Grammar and functions	5 items	Unit 7, lesson B
Task 3	Grammar and functions	Paragraph	Unit 8, lessons A + B
Task 4	Grammar and functions	10 items	Unit 9, lessons A + B
Task 5	Writing	Paragraph	Units 7, 8, and 9 (optional topics)
Task 6	Listening comprehension	Not mentioned	Topic: "homeboys"

According to table 20, tasks one through four focus on the grammar and functions content of units 7 through 9 of the Passages course book, task 5 focuses on the

assessment of the writing skills with topics covered in units 1 through 3, and task six focuses on the assessment of listening comprehension, whose topic is "homeboys". Table 20 also shows that this mid-term test does not assess vocabulary or reading comprehension.

This mid-term test also presents shortcomings regarding the rubrics and the task format in three out of six tasks (see appendix A). Task one, for instance, a mistake correction task, is supposed to assess *relative pronouns in defining relative clauses*, but its format may hinder the testee's performance as the sentences provided are too long and may confuse the test taker. The listening comprehension task (task six) does not present any rubrics or instructions at all. According to the teacher who designed this test, it consisted of a photocopied article (accompanied by a tape with an audio-recorded version of the text) in which some words were randomly deleted using correction fluid. Unfortunately, as in the Adv. 1 mid-term test, the teacher was unable to provide the original text used, with the missing words). Although the lack of this material did not allow me to analyze this listening task, it is safe to assume that gap-filling listening tasks have not been observed in the Passages Two course book. In addition, one may wonder what criterion was adopted for both choosing the text and selecting the words to be erased.

The writing task (task five), or composition, is a three-option task whose topics are valid, but do not exploit the writing processes suggested in the course book. The only exception is option "a", which is the closest one to the writing process suggested in unit nine, "*supporting an opinion or persuasive writing*". As conclusion, it may thus be observed that this test presents the following shortcomings: the formats of four out of six tasks have not been observed in the Passages Two course book, and task rubrics of tasks one, three, and six, might lead to misinterpretation by the testee. In addition, vocabulary and reading comprehension are not assessed.

The absence of a scoring system in this test does not allow for any assumptions regarding the score distribution among test tasks.

Level Adv. 2: final test (Passages two, units 10 through 12): The final test sample selected for the analysis in the present study is composed of four tasks assessing grammar and functions, reading comprehension, and writing (see appendix A). Only task four resembles task formats observed in the Passages Two course book. Table 21 presents information regarding each task in the final test sample designed for level Adv. 2.

Table 21: Level Adv. 2, final test (Passages 2, units 10 through 12): test task construct focus, number of items, and content covered.

Task	Construct focus	# of items	Content or topics covered
Task 1	Presumably grammar and functions	5 items	Topics: Units 10 and 11
Task 2	Grammar and functions	13 items	Past tenses (not part of the content of units 10 through12)
Task 3	Writing	Paragraph	Unit 12 (optional topics)
Task 4	Writing	Paragraph	Unit 10

As can be observed from table 19, tasks one, presumably aims to assess grammar and functions, as none of its task items focus on a specific grammar point or function, or even vocabulary (see appendix A). In this task only the topics of the questions are related to the units this test is supposed to assess (units 4 through 6). Tasks two focuses on a grammar point (*past simple, past continuous, past perfect, and "would")* that is not included in units 10 through 12. Task three and four on the assessment of the writing skills with topics covered in units ten and 12. Table 21 also reveals that none of the grammar and functions content of units 100 through 12 is assessed. In addition, the following constructs are absent: reading and listening comprehension, and vocabulary.

This final test presents shortcomings regarding the rubrics and the task format, as well as the content it aims to assess (see appendix A). In task one, for instance, whose rubrics are "Answer these questions using your own ideas and the number of lines provided", it is difficult to determine what grammar point or function is assessed. It consists of five questions: three of them involve discussion topics covered in the course book's oral activities, while the other two consist of general topics, not included in the content this test is supposed to cover. Only questions "a", "b" and "e" might trigger the

use of the present perfect simple tense, one of the tenses covered in unit eleven-lesson B, but further assumptions would be far fetched. Task two, a gap-filling task, assesses verb tenses (past tenses) that are not part of the grammar content supposed to be assessed (see appendix A). The grammar and functions content of Unit 10, lesson B presents an overview of passives in present, past, and future aspect tenses, but task two concentrates on the assessment of past tenses in the active voice. Moreover, a passive verb construction in task two is possible in only one gap out of thirteen.

In the composition task (task three), although the two topics are valid, neither of them explores the writing processes presented in the units (such as *writing a job-application letter*, for instance). Task four, on the other hand, assesses summary-writing skills, which is the writing process presented and practiced in unit ten. We may thus arrive at the conclusion that except for task four, this final test presents shortcomings in terms of task rubrics and formats, as well as the content it aims to assess. More precisely, apart from the fact that this final test does not assess reading and listening comprehension, or vocabulary, none of the expected grammar and functions content is properly assessed.

Just as in the mid-term test, the absence of a scoring system in this final test does not allow for any assumptions regarding the score distribution among test tasks.

In the following section I will concentrate on the analysis the extent to which the written tests used in the present study contain the characteristics of test usefulness in the model proposed by Bachman and palmer (1996).

4.2.3 The analysis of usefulness

For the analysis of usefulness a parallel has been established between Bachman & Palmer's (1996) model of test usefulness and the outcomes of the written tests described and analyzed in the previous section, more specifically in terms of constructs assessed (grammar and functions, vocabulary, listening comprehension, reading comprehension, and writing skills), test content (more specifically grammar and functions, vocabulary,

and essay topics), test tasks (the extent to which they resemble those of the course book), scoring procedures, and resources for test design and test administration. Each test usefulness quality in Bachman & Palmer (1996) has thus been combined with the above characteristics, as follows:

Reliability: Bachman & Palmer (1996) state that in order to obtain a minimum accepted level of reliability one needs to consider "the purposes for which the test is intended" (p. 135). Each written test is designed to be administered once, in a specific group of students of a particular level, therefore it must be analyzed as a single measuring instrument for a single group of students in a single occasion. In other words, it is not possible to determine a test's level of reliability by comparing it to other tests designed for the same level, assessing the same content. Thereby the analysis of reliability focused on the consistency of measurement across a written test's tasks. More specifically, I investigated whether the language ability components in each test (regardless of whether any of the components supposed to be assessed – listening comprehension, grammar and functions, vocabulary, and writing skills - are absent) are assessed with uniformity along the test, in terms of both quantity (number of tasks and number of items in tasks), and scoring (distribution of points among items and tasks).

Construct validity: this quality is strictly related to what Bachman & Palmer (1966) call "appropriateness of the inferences made about the test taker's language ability" (p. 140). In other words, the analysis of construct validity investigated whether the test assesses what is supposed to be assessed. More specifically, it has been examined whether all supposed content and topics of the related course book syllabus (namely that of listening and reading comprehension, writing skills, grammar and functions, and vocabulary) is assessed. While the analysis of reliability focused on uniformity among test tasks and items, the analysis of construct validity focused on test content. The analysis of construct validity also aimed at investigating whether task rubrics are clear enough and provide examples so as to avoid misinterpretation by the testee.

Authenticity: This quality measures the relation of test task characteristics with the TLU (Target Language Use) domain tasks (Bachman & Palmer, 1996). Thus, in the present study, the analysis of authenticity consisted of determining the nature of the test tasks, more specifically, whether the test tasks resemble those observed in the course book.

Interactiveness: This quality pertains to the involvement of the test taker with the test tasks when performing it (Bachman & Palmer, 1996). The present analysis investigated the extent to which task topics relate to the test taker's own topical knowledge, especially when it requires subjective responses from the candidate. In other words it has been investigated whether there are more contextualized tasks in the test that require the test taker's use of his own personal characteristics or knowledge of the world. Furthermore, it was also been investigated whether there are any tasks that would contain emotionally charged or controversial topics (especially in tasks that require a personal response from the candidate), which would cause the candidate to produce a negative affective response.

Practicality: Bachman and Palmer (1996) state that this quality encompasses 'the amount of resources required and available during the different stages of the test development process" (p. 148): test design, test operationalization, and test administration. For the purposes of the present study, these stages have been simplified to the following stages: (a) test design (the actual "creating the written test" stage), and (b) test administration (the stage in which test takers sit the test, taking into account the following resources: time and space allotment, and additional equipment and materials needed). In the analysis of this usefulness quality it was investigated what these resources are (required and available) specifically when administering the written tests.

As I have stated in chapter one, given the scope of the present study, the sixth quality, *impact* will not be addressed. Measuring this test quality would call for extended research period, as well as other specific instruments, such as class

observation and additional interviews with students, teachers and even course coordinators.

The outcomes of the analysis are presented below under each test quality heading:

Reliability: In terms of reliability it has been observed that in twelve out of twenty tests the components were assessed in an imbalanced way, that is, throughout theses tests, the number of items was different in each test task. This lack of uniformity may lead us to believe that more emphasis is given to some components - grammar and functions, for instance - than to others - listening comprehension or vocabulary, for instance. In the mid-term test for level 5, for instance, task F, which assesses the use of *past continuous* and *past simple,* contains 11 items, while task G, which assesses *past perfect* and *past simple,* contains only 5 items.

The analysis of reliability also pertains to how test tasks and their items are scored, or, in other words, how test takers' responses are quantified in order to arrive at a final score (Bachman & Palmer, 1996). Therefore, the primary objective in this aspect has been to refer to the scoring system each teacher applied in their tests in order to arrive at the final score (100 points, for instance). Twelve out of twenty tests do not present a scoring system, so in these cases the scoring depends virtually on the mind of the teachers who designed them. In those eight tests that do include a scoring system, the distribution of points is not the same for all test tasks. In the final test for level 2, for instance, tasks one, two, three, four, and five are worth 1 out of ten points each, tasks six and eight are worth 1.5 out of ten points, whereas task seven is worth 2 points out of ten. The only exception is the final test for level 1, which consists of a fac-simile copy of test two in the first volume of the New Interchange course book. In the scoring system for this written test, the distribution of points among tasks may be considered as uniform (see description of tests, in section 4.2.1.).

Based on what has been stated above in terms of assessment uniformity of test content, as well as the scoring procedures in the written tests, we may thus conclude

that, except for the final test for level 1, the written tests sample used for the analysis in the present study fail to be reliable.

Construct validity: Unlike reliability, which concentrated on task item and scoring uniformity, the analysis of construct validity consisted of determining whether tasks assess the specific syllabus content supposed to be assessed in each test, more specifically, in terms of listening and reading comprehension, writing skills, grammar and functions, and vocabulary knowledge. The analysis of construct validity also focused on the effectiveness of task rubrics.

None of the twenty tests include the whole syllabus content from the respective units they are supposed to assess (that is, at least one listening comprehension task, one reading comprehension task, tasks assessing all grammar and functions, and vocabulary content, and one writing task). In terms of the assessment of vocabulary knowledge, twelve tests out of twenty do not even assess this construct, and only six tests out of twenty assess part of the vocabulary knowledge to be assessed (see tables in section 4.2.1.). In some tests inconsistencies reached severe levels, such as in the final test of the advanced 2 level, where almost none of the supposed content is assessed (see table for Adv.2, final test, in section 4.2.1.).

Another aspect taken into consideration in the analysis was in terms of task instructions or rubrics. Nineteen out of twenty tests analyzed contain at least some tasks whose rubrics are not clear or lack extended directions. In addition, few task rubrics provided examples, a helpful tool in optimizing test taker performance. The interviews with teachers (see appendix C) revealed why some of them choose not to include one or other component in their tests.

Given the fact that none of the written test samples analyzed in the present study assessed all constructs and syllabus content they were supposed to assess, and as the task rubrics were prone to cause misinterpretation by the testee, the tests yield a low degree of construct validity.

Authenticity: The analysis of authenticity has investigated the extent to which test tasks resemble those of the course book. In all tests the tasks are mainly those requiring *selected responses* - tasks containing items in which there is only one correct answer, such as multiple choice tasks, true-false tasks, among others (Bachman & Palmer, 1996), which commonly assess listening and reading comprehension - and those requiring a subjective answer from the student, or *limited production* tasks (Bachman & Palmer, 1996). In all written test samples analyzed both selected response and limited production types have been observed, assessing all constructs - grammar and functions, vocabulary, and even listening or reading comprehension.

In three out of twenty tests, all tasks resemble those observed in the course book, with only slight adaptations (level 2, mid-term; level 3, mid-term and final tests). In twelve out of twenty tests, around 80% of the tasks resemble those of the course book (level 1, mid-term and final tests, and level 2, mid-term test, for instance), and in only five out of twenty tests the number of tasks that resemble those observed in the course book represent 10 % of all test tasks (level 7, final test, for instance). However, with regards to the writing skill tasks, in five out of twenty tests the writing topic does not relate to the specific course book units' topics or does not fully explore the course book units' central writing process (level 7, final test, for instance).

As the number of tests containing tasks that resemble those observed in the course book is high, (fifteen out of twenty tests in which either all or most tasks resemble tasks in the course book), the written test samples used in the present study may be considered to be highly authentic. Those teachers who used tasks that were adapted from those in the tests suggested in the course book teacher's guide, took the advantage of using authentic tasks. This is explained by the fact that the tests in the teacher's guides are mainly composed of the same task formats and topics as those in the course books.

Interactiveness: The analysis of interactiveness investigated the degree of involvement of the test taker's topical knowledge in performing the test tasks, as well as

the existence of controversial topics that in any way could hinder the test taker's performance. Tasks with a high level of interactiveness are those in which the test taker needs to refer to his or her personal characteristics or knowledge of the world (Bachman and Palmer, 1996). Ten out of twenty tests present a low level of interactiveness, that is, they contain only one or two tasks that require the personal involvement of the test taker (level 4, mid-term test, and level 6, mid-term test, for instance). Only five of all tests present a high level of interactiveness (in which the majority of tasks call for great involvement of the test taker's own experience or topical knowledge (level 7, mid-term test, and level 8, mid-term and final tests, for instance). The remaining five tests yield a medium level of interactiveness, in which at least half of the tasks in each test requires the test taker's personal characteristics or experience (level 4, final test, and level adv. 1, mid-term test, for instance). There were no tasks to be considered offensive or containing potential controversial topics.

Given the number of tests that contain tasks which allow the testee to refer to his or her own knowledge of the world, the written test samples analyzed in the present study may not be considered highly interactive.

Practicality: Practicality pertains to the resources that are required for test design and test and test administration, as well as real availability of these sources for maximizing optimal test performance (Bachamn & Palmer, 1996). Regarding the test design stage, most teachers designed their tests in the computer and used a photocopying facility in the university's Departamento de Língua e Literatura Estrangeiras (DLLE). With regard to test administration, the following resources were considered: time and space allotment and additional equipment needed. In terms of time allotment, written tests are supposed to be administered within one ninety-minute class, and regarding space allotment, these tests are usually administered in the classroom. Therefore, when designing tests, teachers are supposed to consider the length of the tests they design so that test completion within this time period is feasible. It has been observed that all tests analyzed were different in terms of task size and quantity, so we

may thus conclude that teachers have different concepts of test size. As one of the language abilities to be assessed in the construct is listening comprehension, it seems that there are no constraints regarding the equipment required for that purpose: teachers use either the classroom CD player, or even the department's audio laboratory facilities.

It may be thus safe to admit that there may be no constraint in terms of materials and space allotment for the administration of the written test samples. With regard to the time allotted, which is directly related to the size of the tests, in order to investigate whether the test samples may be administered within the usual ninety minutes, they should all undergo what Bachman and Palmer (1996) call a *try-out* stage in order to verify if time estimates are consistent. Therefore, it is possible to come to the conclusion that the written test samples analyzed in the present study may only be considered practical in terms of the following resources available: materials and equipment (paper sheets, photocopying facilities, as well as CD players, used either in class or at the university's audio laboratory), and space allotment (the classroom).

Having addressed the analysis of usefulness of the written test samples used in the present study, I will now turn to the interviews with the teachers who designed these written tests.

4.3. The interviews with the teachers

As I have previously stated, the purpose of the interviews was to substantiate hypotheses regarding the results of the analysis and assert the teachers' views when designing, administering, and scoring the written tests they have designed. It has also been established that, except for the e-mailed interviews, the pre-established questions were amassed in a questionnaire that was actually used to guide the discussions so that their views are more accurately represented (see appendixes II and III).

The questions addressed the teachers' engagement in the design, administration and scoring of the written tests they have made. The discussion on the interviews is organized the following way: the teachers' views are presented under specific

highlighted stages in the testing process as they are addressed in the interviews, namely (a) decisions on test content (grammar, functions, and vocabulary) and topics, (b) specific skills to be assessed (reading comprehension, listening comprehension, and writing or compositional skills), (c) task formats, (d) test length, and (e) the scoring procedure. In order not to reveal each teacher's real identity, each is referred to as teacher A, teacher B, and so on.

a. *Decisions on test content (grammar and functions, and vocabulary) and topics:* all teachers (teachers A, B, C, D, E) agree that the test content and topics are taken from the course book used in class. However, teacher B does not include a specific vocabulary knowledge task, only grammar. Vocabulary, according to this latter teacher, is assessed via the writing task, which is usually in the format of a dialog. Teacher B also states that she does not agree with the way the course book presents the units' topics or grammatical and functional structures. She also supports that a topic or structure is not important because it is in the book, but because it is needed in the students' real life. Teacher F, as well, seems not to use the course book as the only reference when designing tests. She argues that she tries to design tests in which the students are not only tested in terms of grammatical content, but she tends to bring into the testing context the students' reflection over their use of the target language.

b. *Specific skills to be assessed (reading and listening comprehension, and writing skills):* only two teachers intended to assess all three skills (teachers A and E). Teachers B and C admit they assess only listening comprehension and writing skills, while teacher D assesses writing skills only. Teacher E reported she focused on the assessment of reading comprehension and writing skills. Regarding the topics and materials chosen for the assessment of these skills, teacher A chooses the reading and listening topics and materials from other course books or materials whose topics are the same as those of course book used in classroom. Teacher A reports that he or she prefers easier listening tasks. Writing tasks, however, are sometimes taken from the course book used in class. Teacher B uses the recorded material in the teacher's guide and chooses the topics for

writing either according to the course book, or a "reader" that she asks the students to read during the semester. Teacher C also uses the listening material from the teacher's guide in her tests and for the composition she uses the topics from the course book, with slight changes or adaptations. Teacher D also takes the writing topics from the course book, whereas teachers E and F use extra materials whose topics are connected to those covered in course book units. It is thus safe to assume that each teacher follows his or her own criteria when deciding what skills to assess and where to obtain the materials for the design of these tasks.

c. *Task formats:* teachers A and B base their task choice on intuition. Teacher A does not follow any specific criteria, but may sometimes use adapted versions of tasks in the teacher's guide tests. Teacher B prepares writing tasks based on the grammar focus boxes in the course book (which could actually be characterized as a grammar or function task), but tries not to use the same tasks as those in the course book when it comes to assessing grammar, except for a few in the workbook from which she obtains some ideas in the design of alternative tasks formats. Teacher C, based on her claim that the tasks in the course book are too easy and mechanical, also tries to use adapted or altered versions of tasks in the course book, or even creates her own tasks, whereas teacher D tries to devise test tasks based exclusively on those of the course book. Teacher E claims that some of the tasks he devises stem from those formats observed in the course book, though most tasks are created using extra sources, such as other course books or the Internet. Teacher F also reports she prefers to create her own tasks based on materials other than the course book.

d. *Test length:* teacher A admits not to think about test length, but states that devises her tests so that they can be accomplished in the period of one hour and thirty minutes. Teacher B also shows her concern regarding the time allotted for each of the test tasks. Teacher C recognizes that experience has shown her that test and task length had to be reduced so that test takers could perform all tasks within the allotted time (one hour and thirty minutes). Teacher D also shows concern regarding the number of test

pages and task complexity and admits that she bases her test size assumptions on her teaching experience. Teacher E states that the size of his tests is determined by how much a good student is able to perform in one hour, whereas teacher F considers the amount of time all students would need to perform the test. Although all teachers take test length into consideration, one may come to the conclusion that the assumptions regarding this aspect are based on intuition or experience.

e. *Scoring procedure:* teacher A admits she makes occasional use of the scoring suggestions in the teacher's guide tests. Although tasks may have different numbers of items, teacher A tries to give all tasks the same weight. Teacher B, on the other hand, usually gives a higher score for listening comprehension and the composition (thirty percent for each, out of the total score). The remaining forty percent are distributed among four remaining tasks assessing grammar and functions. Teacher C also adopts a similar procedure by distributing forty percent of the total score to both the listening comprehension and the writing skill tasks respectively; the other sixty percent are distributed among the remaining tasks assessing grammar, vocabulary knowledge, and functions. However, she admits more weight is placed on more complex tasks. Teachers D and E also report that in their tests more complex tasks are worth a higher score teachers, whereas in teacher F's tests all tasks have the same weight in scoring, except for the composition (writing skill), which is usually worth twenty percent of the total score. In addition, both teachers D and F claim that at times a certain percentage of the total score in assessment is devoted to student development and participation in class. Regarding the scoring of individual items inside test tasks, teachers adopt different procedures as well. Teachers A, B, and C, for instance, claim that in items where grammar accuracy is being assessed, half-correct or partial credit scoring may be considered, whereas in reading or listening comprehension items, only the correct factual information is considered. Teacher D does not report her procedures regarding the scoring of items that require a subjective answer from the test taker, and teacher E, in the same situation, admits she uses five different levels of correction: 100% correct,

75% correct, 50% correct, 25% correct, and 0% correct. Regarding scoring procedures for the writing or compositional skills, most teachers report to base their scoring on aspects such as textual cohesion and coherence, appropriateness of vocabulary, grammar accuracy, and spelling. However, none of them were able to give details about the rating of these different writing ability components.

As can be learned from the teacher's views expressed in the interviews, although there is a significant reliance on the course book used in class, which explains the level of authenticity of the tasks in the written test samples used in the present study, for instance, on the whole, teachers tend to base test design and scoring on their own intuition and teaching experience. In the next section, I will discuss the results of the test usefulness analysis carried out in the present study, as well as the teacher's views revealed by means of the interviews.

4.4. The discussion of results

The analysis of *reliability* has revealed that in twelve out of twenty tests the language ability components or constructs were assessed in an imbalanced way, that is, some tasks were weighted both in terms of number of items assessing a specific grammar and functions content, for instance, or in terms of scoring, with some tasks being worth more than others.

In the few tests (four tests out of twenty) that assess all constructs, more emphasis is given to some components at the expense of others. Assuming that in the TLU domain (the course book syllabus) the five language ability components are practiced in a uniform manner, that is, grammar and functions, vocabulary knowledge, listening comprehension, reading comprehension, and the writing skills are given equal emphasis in the instructional program, it is expected thus that the same procedure should be adopted when assessing these components in a test. The interviews, however, revealed that some teachers, based on their own beliefs and assumptions, deliberately either omit

some of these components, or place more emphasis on some tasks at the expense of others (see appendix C).

Teacher B, for instance, does not assess reading comprehension in her written tests, but encourages extra class reading by means of extra readers, which is used at the end of the semester to assess reading comprehension by means of the oral test, thus integrating both reading and speaking skills. Teacher B does not include a specific task assessing vocabulary as she claims that vocabulary knowledge is assessed via the writing task, or composition. Teacher D does not include reading or listening comprehension tasks in her tests, claiming that, in the case of the former, test takers take too much time performing the tasks, and as for the latter, since there is plenty of listening practice in the classroom, teacher D argues that it would not be necessarily appropriate to include a listening comprehension task in a written test. Teacher E does not specify why he does not include a listening comprehension task in his written tests.

In terms of scoring, the analysis revealed that, as in most tests the score distribution among tasks and task items is not uniform, reliability is not consistent. Moreover, as most tests do not present their scoring system written on them, it is not possible to establish any assumptions regarding the distribution of points. However, the interviews unmasked a few procedures teachers adopt (see appendix C). When enquired about test scoring, teacher A reports that she tries to give all tasks the same weight. However, all other teachers claims that they place more weight in certain tasks than in others, either in the listening comprehension or compositional tasks, or in more complex grammar and functions tasks. When scoring individual items, however, all teachers state that in tasks where comprehension is elicited (as in listening or reading comprehension tasks) grammar accuracy and spelling do not count in the scoring, but in the case of composition scoring, the interviews confirm that teachers do not use a specific rating scale. What actually seems to occur is that teachers arrive at a single score based on their intuitive rating of different writing constructs.

The analysis also reveals inconsistencies with regard to *construct validity.* Although five out of six teachers confirm that they based the assessment of content and topics on the course book (see appendix C), in most tests there is a lack of two or more components of either grammar or functions, or vocabulary content supposed to be assessed. In other words, teachers deliberately left out the assessment of certain grammatical and functional structures, or vocabulary covered in the course book units. I observed that task rubrics in nineteen tests out of twenty were not clear or lacked extended directions or examples, and only teacher C admits to have improved task instructions since the semester she designed the test sample used for the present study. In addition, teacher C claims that before test takers begin the tests, she orally checks that they understand what they are supposed to do.

In terms of *authenticity*, it has been observed through the analysis that, although there have been some modifications, most teachers devised test tasks that resembled those of the course book, even if in the interviews some report their preference to use other sources for that purpose (teachers A, C, E, and F, see appendix C). Their views concerning the choice of test tasks differ according to their own pedagogical beliefs. Some teachers (teacher B and C, for instance) claim that the course book task formats were not appropriate to be used in tests, whether students had a negative reaction towards them (teacher C), or because they are not similar to those in real life (teacher B). Although according to the analysis the level of authenticity may be considered relatively high in most tasks assessing grammar and functions, vocabulary, as well as reading and listening comprehension, the choice of the writing tasks still allows for inconsistencies: in five out of twenty tests these tasks do not specifically relate to the course book's writing approach, and are thus not representative of the writing tasks in the TLU domain. Teacher B, for instance, in the test sample analyzed in the present study, created her own writing tasks, in which the test taker had to write a dialog using the grammatical structures taught in the course book. One may thus wonder how authentic a dialog-writing task may be if compared to real life writing tasks.

Another important aspect that has emerged with the analysis of *authenticity* pertains to the tests suggested in the New Interchange series' teacher's guide. It should not come as a surprise that some tasks in the written tests designed by teachers resemble those tasks in the written tests suggested in the course book's teacher gui des. In some extreme cases the whole test was simply cut and pasted, with only slight changes so as not to look one hundred percent similar. I have observed that most tasks in the teachers' guide tests seem authentic as their formats and characteristics are almost copies of those in the series' student's book and workbook. Although teachers are expected not to use these tests, this practice is very common as it saves time – a claim commonly heard while I informally chatted with them. However, a point that must be taken into consideration is that by referring to these tests teachers actually make use of authentic material specifically designed to assess the construct of the course book's syllabus.

In terms of *interactiveness*, although only some tests include tasks that require the test taker's personal involvement, there is among teachers a certain concern regarding an affective reaction to the test tasks (see appendix C). In the writing task, teacher A, for instance, claims that she encourages test takers to write about a subject they are familiar with. However, although not much was stated in the interviews in terms of the involvement of the test takers' own topical knowledge, it was observed that a few teachers do try to elicit the test taker's knowledge of the world when performing some tasks. This is evident in tests such as the mid-term test of level 7 and both mid-term and final tests of level 8, for instance (see appendix A). It may thus be observed that nine tests out of twenty may be considered as containing interactive tasks.

The analysis of *practicality* has looked into aspects of test design and test administration. Regarding test design, since the tests analyzed are written tests, no constraints have been observed. In terms of administration, however, time allotment is a variable that may be affected by the length of the tests. As I have mentioned in the analysis, tests are of different sizes. As test takers sit the test for the period of one hour and thirty minutes, a long test may negatively affect their performance. In the interviews

(see appendix C) teachers reveal that they base test length on their own intuition and testing experience.

Each test is rarely administered more than once, as students are allowed to keep their own copies of them. In other words, as a result, a specific written test is always devised for a specific group of students. This means that it is difficult to anticipate how much time will be needed for test takers to perform the whole test. Therefore all tests analyzed are prone to have been problematic in terms of time allotment when they were used. In other words, in terms of time allotment, as Bachman & Palmer (1996) advocate, a test is only possible to be considered practical if it goes through a trial stage. It is thus safe to assume that in order to consider these tests practical in terms of allotted time, further evidence is needed by having a group of test takers sit each test at least one more time.

Having discussed the outcomes of test usefulness analysis, as well as the teacher's views in order to confirm the hypotheses that stemmed from the present study, I will now address the research questions that have stimulated my engagement in the present study. The answers are provided immediately after each question:

1. Do the achievement tests contain the following usefulness qualities, namely *reliability, construct validity, authenticity, interactiveness* and *practicality,* as proposed in the Bachman and Palmer (1996) model? It has been observed that none of the tests do contain all qualities in one test. In some tests certain qualities may be more salient at the expense of others. There are tests, for instance, whose tasks may be considered more interactive than those in other tests, but are not necessarily as authentic. In other words, in the attempt to design a more interactive task, a teacher might choose a task format which does not resemble any of those practiced in the classroom beforehand, thus running the risk of minimizing optimum performance by the testees as they might not feel familiar with the test task.

Reliability is a usefulness quality that has not been observed in the majority of tests. The analysis of usefulness revealed that rating, for instance, is prone to

subjectiveness. In other words, as revealed in the interviews, teachers use their own intuitive criteria when correcting the test compositions, for instance. This may lead to rating and scoring inconsistencies across different compositions. The same measurement shortcoming might occur while rating and scoring items assessing grammatical content, such as *limited response* items – those that require a subjective answer by the testee. The final test of level 8, for instance, contains several *limited response* items in which different responses by the different testees might not be reliably rated and scored, since students' responses will all vary in terms of amount of information provided.

In terms of *construct validity*, the analysis revealed that none of the tests may be considered 100% valid, as all of them lack one or more constructs, or part of the content supposed to be assessed, but the analysis of *authenticity,* on the other hand, brought evidence that the great majority of tests (fifteen out of twenty) contain tasks that resemble those observed in the course book and workbook of both series "New Interchange" and "Passages", which means that their characteristics seem to correspond to the Target Language Use (TLU) domain at a large extent. However, another fact that cannot be discarded is that teachers may use extra material in class, collected from other sources, such as other course books, reference books for extended practice of skills, such as reading and listening comprehension, and writing, and also from the internet. Therefore, we may admit that the extent to which certain test tasks resemble these extra tasks other than those of the course book must be considered as well. More specifically, as each test is designed to be administered only once in one particular group of students, the extent of task authenticity, in its specific situation, may even be larger. However, this is a variable that calls for further observations, and will be discussed in the next chapter in the section that addresses the limitations of this study. In addition, although all tests may bear a high level of *practicality* in terms of materials and equipment available, further research is needed in order to investigate whether it is possible for testees to perform the test tasks within the amount of time available.

As it has been observed that different tests contain different usefulness qualities at different extents, supporting Bachman and Palmer (1996), it may thus be concluded that the written tests analyzed cannot be considered balanced or uniform in terms of test usefulness qualities.

2. How do teachers design the written tests for the EFL extension program at UFSC ? The interviews with the teachers consisted of a valuable source of details regarding how teachers design their tests. Although most teachers agreed that the test content in based on the course book syllabus, their answers to the questions revealed that the choice of content and specific skills to be assessed in the tests, as well as the task formats, and scoring procedures, all seem to depend almost exclusively on their own subjective criteria and their personal concepts of what language assessment is about. The interviews lead us to the conclusion that, given the facts reported by the teachers, language assessment in the EFL extra curricular program at UFSC seems to be entirely dependent on its teacher's own intuition and beliefs, which might have developed from past experiences these teachers had in the roles as both students and teachers.

In the next chapter, the conclusion, I will present a summary of the results obtained from the test usefulness analysis in the present study, as well as the limitations of the study and provide insights for further research and pedagogical implications.

CHAPTER V

CONCLUSION

In this section the following issues are covered: first a summary of results of this study is presented, and its limitations are discussed. Finally, insights for further research and pedagogical implications are addressed.

5.1. Summary of the study

The main aim of the present study was to carry out an analysis of the usefulness of the achievement written tests (mid-term and final tests) designed by teachers in the EFL extension program at Universidade Federal de Santa Catarina (UFSC). The analysis carried out was based on the framework proposed by Bachman and Palmer (1996). In order to complement the findings that stemmed from the analysis, interviews with teachers were carried out by means of a pre-established questionnaire, which was used in order to elicit the teachers' views regarding the stages of test design and scoring methods. For the purpose of the analysis five of the six test usefulness qualities were addressed: *reliability, construct validity, authenticity, interactiveness,* and *practicality.* The sixth quality, *impact,* was not addressed, as it would require specific instruments to be measured for the scope of the present study.

The analysis of *reliability* investigated whether the language ability components in each test are assessed with uniformity along the test, with respect to the quantity of tasks and task items, as well as the distribution of points among items and tasks (scoring system). In the analysis of *construct validity* I examined whether all supposed content and topics of the related course book syllabus is assessed in each test, as well as whether task rubrics are clear enough and provide examples in order to avoid misinterpretation by the testee. The analysis of *authenticity* consisted of investigating whether the test

tasks resemble those observed in the course book, and the analysis of *interactiveness* investigated the extent to which task topics relate to the test taker's own topical knowledge or knowledge of the world, especially when it requires subjective responses from the candidate, and also whether there were any tasks that would contain emotionally charged or controversial topics. Finally, with regard to *practicality,* I investigated whether required resources concerning test design and test administration were available.

The analysis of usefulness yielded the following findings: tests contained a low degree of *reliability* as in twelve tests out of twenty, their components were assessed in an imbalanced way (that is, the number of items was different in each test task), thus giving more emphasis to the assessment of some components at the expense of others. Regarding scoring, tests also failed to be reliable as twelve out of twenty tests did not contain a scoring system written on them, and the remaining ones that did include a scoring system, the distribution of points among tasks was uneven.

Tests also failed to yield a satisfactory level of *construct validity* for two main reasons: firstly, because none of them assess the whole syllabus content and constructs from the respective units they are supposed to assess, and secondly, due to the fact that nineteen out of twenty tests analyzed contain at least some tasks whose rubrics are not clear or lack extended directions or examples.

Despite the low levels of *reliability* and *construct validity,* the written tests samples analyzed in the present study seem to yield a high level of *authenticity*, as fifteen out of twenty tests contain tasks that resemble those observed in the course book.

On the whole, the written test samples were not considered very interactive. Only five of all twenty tests present a high level of *interactiveness*, as the majority of their tasks allow great involvement of the test taker's own experience or topical knowledge. Although there were no tasks to be considered offensive or containing potential controversial topics, half or less than half of the tasks in the remaining fifteen tests

require the test taker to perform by referring to their personal characteristics or experience.

All written tests yield high levels of *practicality* since resources such as materials and equipment (paper sheets, photocopying facilities, as well as CD players), and space allotment (the classroom) are indeed available. However, it was not possible to determine if tests are practical in terms of time allotment, as this would require all tests to be administered again in order to verify the consistency of time estimates by the teachers.

The interviews with the teachers of the EFL extension program at UFSC revealed that despite their reliance on the course book used in class, teachers tend to base test design and scoring on their own intuition and teaching experience. Among the findings it should be highlighted that some teachers deliberately either omit some of these components, or place more emphasis on some tasks at the expense of others. In addition two of the teachers also admitted their preference for alternative task formats and criticized task formats in the course book by saying that either they did not carry any resemblance with tasks in a real-life context, or by claiming that students were dissatisfied with them. In sum, the main aspect unveiled by the interviews is the fact that teachers base their design of written tests on their own intuition and their teaching and testing experience gained throughout their careers.

5.2. Limitations of the study and further research

What the findings of this present study mainly suggest is that teachers refer to their own conceptions and intuition when they assess and evaluate their students. Despite the concrete findings the present study produced in terms of the usefulness of the written test samples, some limitations need to be taken into consideration. Further assumptions regarding the study findings were posited as insights for forthcoming research, stimulated by each limitation described below.

The first limitation refers to the data collection stage. Ideally, as the number of levels in the EFL extension program at UFSC is ten (levels 1 through eight, Advanced 1 and Advanced 2) the test samples used for the present study should have been those designed by ten different teachers (a different teacher for each level). However, as the number of groups of basic levels (levels 1 through 4, for instance) is larger than the number of groups in late basic, intermediate, and advanced levels, the option of teachers who taught these first levels was larger than that of more advanced levels. Not all teachers wanted to participate in the study and difficulties were found to contact the teachers for level 4 and Advanced 2, for instance. As a result, as teachers are allowed to teach different levels in the same semester, in order to obtain written test samples of all levels, the mid-term and final test samples for level 4 had to be those designed by the same teacher who designed both tests for level 3, and test samples designed for Advanced 2 used in the present study were those designed by the same teacher who designed the ones for Advanced 1. If more teachers had actually volunteered, the data for the present study would have consisted of ten pairs of mid-term and final tests designed by ten different teachers, and thereby the analysis of the present study might have yielded more accurate findings. Thus, for further research it is suggested that a greater range of teachers and tests are used as data.

Another limitation has been observed during the stage of the interviews. Ideally all interviews should have been carried out and audio recorded in person (teacher and researcher). Unfortunately, despite all efforts, the teacher for level 6, for instance, did not reply to the requests for the interview, and two other teachers (teacher E and teacher F) agreed to answer the questionnaire via e-mail only. If all teachers had agreed to be interviewed orally and audio recorded, further valuable details and facets regarding test design and administration might have been revealed. Further research, therefore, would include a larger number of teachers to be interviewed.

A third limitation pertains to the analysis of usefulness using Bachman and Palmer's (1996) framework. Bachman and Palmer (1996) claim that no test is useful for any situation or context, and there is no "best" test. A certain test may be useful in a particular context or for a particular group of students, but it may be inadequate for other groups of students or contexts. This situation seems to be applicable in the context of the present study (the EFL extension program at UFSC), as each test is designed to be administered once. In other words, the teacher of a particular group has the advantage of being able to design both mid-term and final tests whose specifications are exclusively aimed at that specific group of students. Therefore the test samples used for the present study cannot be considered an all-time representative of the entire EFL extension program as they were designed bearing in mind a specific group of students in a specific moment or stage of their course. The extent to which these tests were analyzed and considered useful under each of the usefulness qualities taken from Bachman and Palmer's (1996) model corresponds to the unique time they were designed and administered. It would be sensible to suggest that a further study could be carried out in a broader scope, using test samples designed and administered in different semesters (the first semester of 2003, for instance) by even different teachers, the analysis would probably yield more accurate outcomes.

The fourth, and perhaps one of the most significant limitation, still regarding Bachman and Palmer's (1996) model of test usefulness qualities, involves a sixth test quality of the model, namely *impact*. This quality was not included in the present study, as measuring it would call for special instruments (such as specific questionnaires to both students and teachers), as well as extra time in order to carry out extended interviews with teachers and students. Bachman and Palmer (1996) define *impact* as the many ways a test may affect an educational system or society – the "macro" level – and the individuals directly involved in the test taking experience: the students and the teachers – the "micro" level. It is thus my own belief that measuring *impact* alone would

consist of a separate study per se. If the main instrument of the present analysis was the set of written test samples and their usefulness, a second study would embrace a broader universe consisted of those involved in the test taking experience. Engaging in such study would require the researcher to establish a close contact with those who take the test (the students) and those who design them (the teachers) – the "micro" level, as Bachman and Palmer (1996, p. 29-30) suggest. Regarding how students are affected by taking tests, interviews must be carried out investigating aspects of test taking experience and preparation for it, the feedback received about their performance, and their awareness of decisions made about their test scores (Bachman and Palmer, 1996, p. 31). With respect to how teachers are affected by tests, another set of interviews would yield information regarding the positive or negative *washback* effect in the program. More specifically, following suggestions that Bailey (1996) poses, the study could investigate whether teachers are aware of the real purpose of the tests they design, the need for clear and interpretable test results, and whether these results are fair and credible to their students.

5.3. Pedagogical implications

It is my belief that the present study as well as further ones suggested in the previous section will allow teachers and students to better understand the connection of test taking experience with the teaching and learning practice.

Language testing practice in the EFL extension program at UFSC lacks standard-based procedures in terms of written test design, that is, every time a written test is to be administered, each teacher is supposed to design a test for immediate use. As previously stated, these tests are not supposed to be used more than once, since students are allowed to keep their own tests duly corrected and scored by their teachers. Thereby, the EFL extension program does not make use of pre-designed tests, that is, tests designed to be continuously used throughout the semesters, a common practice among certain

EFL institutes in Brazil. If, on the one hand, the use of pre-designed tests would be a step towards testing practice standardization in the program, teachers, on the other hand, might base their classroom practice on the tests that follow, an attitude called 'teaching to the test" (Bachman and Palmer, 1996, p. 33), thus promoting negative washback, according to Bachman and Palmer (1996). The main advantage of having teachers design their own tests is that these are prone to be taylor-made, or designed for their specific purpose, in a specific group of students, thus increasing their level of usefulness, as Bachman and Palmer (1996) advocate. If the analysis of usefulness of the written test samples used for the present study indicated that teachers are not aware of any current theory and practice of language testing by using their own intuition and personal beliefs to design and administer written tests, then some light might be shed upon developing a specific training program for the teachers in the context of the present study.

This training could consist of a series of workshops composed of the four different sessions: the first session would concentrate on theoretical issues and would aim at making teachers aware of the basic principles that underlie language testing. A second session would deal with the analysis of usefulness of previously designed tests. In other words, participants of the workshop (the teachers) would be given tests and the trainer would ask them to analyze their usefulness by means of a simplified version of Bachman and Palmer's (1996) model of usefulness. In a third stage these teachers would be required to put into practice the input received in the two first sessions by designing a model of written test based on the syllabus of the course book used in the program. Finally, in the fourth session, teachers would present their tests by means of short seminars to the other participants and the tests' usefulness would thus be discussed. The material used in the training program would be extracted of available literature in language testing, such as Heaton (1975; 1988), Hughes (1989), Weir

(1993), Alderson, Clapham, and Wall (1995), Genesee and Upshur (1996), Bachman and Palmer (1996), and McNamara (2000).

It is thus hoped that the present study may stimulate further research not only in its context, but also in other EFL extension programs in Brazilian universities. The above described training program on language testing for teachers, conceived after the results of the present study, could bring immediate rewards, allowing teachers to reflect on their teaching and testing practice, thus establishing a starting point in search for more standardized and desirable quality test design and administration in the EFL extension program at UFSC.

REFERENCES

Alderson, J.C., Clapham,C., & Wall, D. (1995). Language Test Construction and Evaluation. Cambridge: C.U.P.

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

Bachman, L & Palmer, A. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: O.U.P.

Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. Language Testing 2000 17 (1), 1-42.

Bailey, K. (1996). Working for washback: a review of the washback concept in language testing. Language Testing 1996 13 (3), 257-279.

Beglar and Hunt (1999). Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. Language Testing 1999 16 (2), 131-162.

Brown, J.D. & Hudson, T. (1998). The alternatives in language assessment. Tesol Quarterly 1998 32 (4), 653-675.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. Language Testing, 1997 14 (1), 3-22.

Davies, A. (1997). Introduction: the limits of ethics in language testing. Language Testing 1997 14 (3), 325-241.

Davies, A. (1997a). Demands of being professional in language testing. Language Testing (1997) 14 (3), 328-339.

Davis, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). Studies in language testing: Dictionary of language testing. Cambridge: University of Cambridge Local Examinations Syndicate.

Fortkamp, M. B. M. (2000). Working Memory Capacity and L2 Speech Production: An exploratory study. Unpublished doctoral dissertation. Florianópolis: UFSC.

Genesse, F. & Upshur, J. (1996). Classroom Evaluation in Second Language Education. New York: C.U.P.

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

Heaton, J.B. (1975; 1988). Writing English language tests. Essex: Longman Group UK Limited.

Hughes, A. (1989). Testing for language teachers. Cambridge: C.U.P.

Laufer and Nation (1999). A vocabulary-size test of controlled productive ability. Language Testing 1999 16 (1), 33-51.

McNamara, T. (2000). Language testing. Oxford: O.U.P.

Paiva, M. da Graça G., Brugalli, M. (Eds.). (2000). Avaliação: novas tendências, novos paradigmas. Porto Alegre: Mercado Aberto.

Richards, J. C., Hull, J., and Proctor, S. (1997). New Interchange: English for international communication. (Vols. 1- 3). Cambridge: C.U.P.

Richards, J. C., Sandy, C. (1999).Passages: an upper-level multi-skills course. (Vol. 1 and 2). Cambridge: C.U.P.

Schmitt, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. Language Testing 1999 16 (2), 189-216.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: washback effect over time. Language Testing 1997 13 (3), 298-317.

Spolsky, B. (1997). The ethics of gatekeeping tests: what have we learned in a hundred years? Language Testing 1997 14 (3), 242-247.

Weir, C. (1993). Understanding and developing language tests. Hertfordshire: Prentice Hall International (UK) Ltd.

Wolter, B. (2002). Assessing proficiency through word associations: is there still hope? System 2002 30, 315-329.