

UNIVERSIDADE FEDERAL DE SANTA CATARINA
Pós-Graduação em Letras/Inglês e Literatura Correspondente

ASSESSMENT OF READING IN ENGLISH AS A FOREIGN
LANGUAGE: INVESTIGATING THE DEFENSIBILITY
OF TEST ITEMS

Celso Henrique Soufen Tumolo

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

DOUTOR EM LETRAS

FLORIANÓPOLIS

Março 2005

Esta Tese foi julgada adequada e aprovada em sua forma final pelo
Programa de Pós-Graduação em Inglês para a obtenção do grau de

DOUTOR EM LETRAS

Opção Língua Inglesa e Lingüística Aplicada ao Inglês

Dra. Mailce Borges Mota Fortkamp
Coordenadora

Dra. Lêda Maria Braga Tomitch
Orientador

Banca examinadora

Dra. Lêda Maria Braga Tomitch (orientadora)

Dra. Laura Stella Miccoli (examinadora)

Dra. Mailce Borges Mota Fortkamp (examinadora)

Dra. Maria da Graça Paiva (examinadora)

Dra. Sonia Gomes Ferreira (examinadora)

Florianópolis, 28 de março de 2005

Para o Adir
a Bel
a Ligia
o Lu
e o Paulinho

Agradecimentos

À Universidade Federal de Santa Catarina
Ao Curso de Pós-Graduação em Letras/Inglês

À CAPES, pelo pagamento de mensalidades de bolsa de doutorado no Brasil, como também pelo pagamento da bolsa do Programa de Doutorado no País com Estágio no Exterior - PDEE

Ao CNPq, pelo pagamento de mensalidades de bolsa de doutorado no Brasil.

À banca examinadora

Agradecimento Especial

À Professora Lêda Maria Braga Tomitch, minha orientadora, pelo acompanhamento em minha trajetória acadêmica, pelos vários cursos relevantes para minha formação e futura docência, pelo apoio incondicional a minha formação externa, pela orientação deste trabalho, e pelo sempre presente sorriso.

Ao professor Fred Davidson, meu orientador externo, que possibilitou minha formação externa na University of Illinois at Urbana Champaign, USA, me orientou na área de testes, em particular nas definições atuais de validade e validação, e pelo colega que se tornou.

Florianópolis, março de 2005

Abstract

Assessment of Reading in English as a Foreign Language: Investigating the Defensibility of Test Items

Celso Henrique Soufen Tumolo

Universidade Federal de Santa Catarina
2005

Profa. Dra. Lêda Maria Braga Tomitch
Supervising Professor

In the present study, the defensibility of test items is investigated in three different testing situations: proficiency tests, classroom tests, and university entrance examinations. The defensibility is judged using the concept of validity as put forth by Messick (1989) for educational testing and by Bachman (1990) for language testing, i.e., in terms of validity of the interpretations and actions based on test item responses, considering the justifications coming from evidential basis and consequential basis. For the investigation, in terms of evidential basis, for construct-related evidence, constructs of language ability and reading ability are used, and for criterion-related evidence, the criterion defined for university studies by Weir, Huizhong, and Yan (2000) and the framework proposed by Bachman (1990) for the assessment of authenticity of test tasks in relation to the criterion tasks are used. In terms of consequential basis, an appraisal of the consequences is based on Bachman's (1990) notion of fairness and Shohamy's (2001) critical perspective of the use of tests. The method used for the investigation is based on recent notion of validity as argument-based proposed by many scholars in the area of testing. It is also based on the validity table proposed by Chapelle (1994) for considerations of the forces of the arguments, both in favor, or against the validity of the interpretation of ability based on the items, with the additional feature of a third column, with the refutation of the argument against, as suggested by Davidson (personal communication, 2004). Considering the arguments provided for each item, it is possible to conclude that some items are defensible and some are not. Some

defensible items focus on skills such as identification of syntax and cohesion, in particular lexical cohesion, inference of word meaning, elementary and propositional inferences, and identification of functional value. Some non-defensible items have the construct-irrelevant aspects of assessing constructs other than reading ability, such as vocabulary knowledge, background knowledge, writing ability, illustration comprehension, and the criterion-irrelevant aspect of assessing tasks not present in the criterion, such as the tasks specific for reading non-academic texts as poems, narratives, advertisements, and technical problems impeding test takers to perform at their level of ability. Other non-defensible items have the consequence of being biased once specific background knowledge is presupposed for their answers. The political and pedagogical implications of the conclusions claim for the choice or development of items incorporating features of validity, in all the facets, considering both evidential basis and consequential basis, so as to make them more defensible.

Number of pages: 246

Number of words: 66.285

Resumo

Assessment of Reading in English as a Foreign Language: Investigating the Defensibility of Test Items

Celso Henrique Soufen Tumolo

Universidade Federal de Santa Catarina
2005

Profa. Dra. Lêda Maria Braga Tomitch
Orientadora

Neste estudo, a defensibilidade de itens de testes é investigada em três situações de teste: testes de proficiência, testes de sala de aula, e vestibulares. A defensibilidade é julgada recorrendo-se ao conceito de validade proposto por Messick (1989) para testes em educação e por Bachman (1990) para testes em línguas, isto é, em termos da validade das interpretações e ações baseadas nas respostas aos itens de testes, considerando as justificativas oriundas das análises de evidências e de conseqüências. Para a investigação de evidências, para evidências relativas ao construto, construtos de habilidade lingüística e de habilidade de leitura são usados, e para evidências relativas ao critério, o critério definido para estudos acadêmicos proposto por Weir, Huizhong, e Yan (2000) e o arcabouço proposto por Bachman (1990) para a avaliação da autenticidade de tarefas presentes nos testes em relação às tarefas exigidas para estudos acadêmicos são usados. Para a investigação das conseqüências, uma apreciação das conseqüências é calcada na noção de justo de Bachman (1990) e na perspectiva crítica do uso de testes de Shohamy (2001). O método usado para a investigação é pautado na noção recente de validade baseada em argumentos, proposta por vários pesquisadores na área de avaliação. É também pautado na tabela de validade proposta por Chapelle (1994) para considerações das forças dos argumentos, tanto a favor como contra a interpretação de habilidade baseada nos itens, com uma terceira coluna adicional com a refutação do argumento contra, como sugerido por Davidson (pessoalmente, 2004). Considerando os argumentos para cada item, é possível concluir que há itens

defensáveis e itens não defensáveis. Alguns dos itens defensáveis focam em habilidade como identificação de elementos gramaticais e de coesão, inferência de significado de palavras desconhecidas, inferências elementares e proposicionais, e identificação do valor funcional do uso da língua. Alguns itens não-defensáveis têm problemas técnicos, como também aspectos irrelevantes ao construto de leitura de avaliar outros construtos, como conhecimento de vocabulário, conhecimento de mundo, habilidade de escrita, compreensão de ilustrações, como também aspectos irrelevantes ao critério de avaliar tarefas não exigidas para estudos acadêmicos, como tarefas específicas para leitura de textos não-acadêmicos como poemas, narrativas, propaganda comercial, todos dificultando que o desempenho nos testes reflita a habilidade avaliada e auxilie na avaliação de desempenho futuro. Outros itens não-defensáveis têm conseqüências de favorecer alguns grupos em detrimento de outros, já que conhecimento específico é pressuposto para algumas de suas respostas. As implicações pedagógicas e políticas das conclusões são que a escolha e desenvolvimento de itens de testes incorporem noções de validade, em todas as suas facetas, considerando tanto a validação de construto, como também autenticidade das tarefas e impacto do uso do teste na sociedade e nos indivíduos, de tal forma a tornar os itens usados mais defensáveis.

Número de páginas: 246

Número de palavras: 66.285

TABLE OF CONTENTS

LIST OF TABLES	xii
LIST OF APPENDICES	xvi
CHAPTER I – INTRODUCTION	01
1.1 – Introduction	01
1.2 – Research questions	05
1.3 – Motivation for the study	06
1.4 – Significance of the study	08
1.5 – Organization of the dissertation	09
CHAPTER II – CONCEPTIONS OF VALIDITY IN TESTING	11
2.1 – Validity in testing	11
2.1.1 – Historical perspective of validity: the evolution of the concept of validity	11
2.1.2 – Validity as a unitary concept: the progressive matrix proposed by Messick	16
2.2 – Validity in language testing	19
2.2.1 – Bachman (1990) interpreting validity to language testing	20
2.2.2 – Alderson, Clapham and Wall (1995) interpreting validity to language testing	27
2.2.3 – Bachman and Palmer’s (1996) notion of usefulness and framework for task characteristics	29
2.3 – Consequential validity to the extreme	36

CHAPTER III – DEFINING THE CONSTRUCTS FOR THE INVESTIGATION OF VALIDITY	44
3.1 – Framework by Bachman: an explanatory approach	45
3.2 – Reading ability: accounts within the explanatory approach	52
3.3 – Defining the scope of the construct to be used for analysis	60
3.4 – Discussing the correspondence of item used skill assessed	65
 CHAPTER IV – THE STUDY	 74
4.1 – Validation and source of validity evidence	75
4.2 – Methods of test validation: arguments and justifications	78
4.3 – Articulating methods to sources of evidence	83
4.4 – Defining the sources of evidence	85
4.5 – Defining the focus of the analysis: relevance and representativeness	90
4.6 – Defining the criterion for the analysis of the university entrance examinations	94
4.7 – Defining the facets within the framework for the analysis of task characteristics for the criterion-related evidence	99
4.8 – Reverse engineering: the approach for collecting validity evidence	111
4.9 – Material used for data analysis	113
 CHAPTER V – INVESTIGATING THE DEFENSIBILITY OF ITEMS IN PROFICIENCY TESTS AND EAP TESTS: EVIDENTIAL BASIS	 115
5.1 – Analysis of the IELTS test items	115
5.2 – Analysis of the TOEFL test items	130
5.3 – Analysis of the items used by EAP teacher 1	154
5.4 – Analysis of the items used by EAP teacher 2	161

CHAPTER VI – INVESTIGATING THE DEFENSIBILITY OF ITEMS WITHIN THE UNIVERSITY ENTRANCE EXAMINATIONS: EVIDENTIAL BASIS AND CONSEQUENTIAL BASIS	174
6.1 – UFSC entrance examinations: evidential basis	175
6.2 – UNICAMP entrance examinations: evidential basis	194
6.3 – Consequential validity: considerations for entrance examinations	225
CHAPTER VII – CONCLUSIONS, FINAL REMARKS, LIMITATIONS, AND IMPLICATIONS	230
7.1 – Conclusions	230
7.2 – Final remarks	233
7.3 – Limitations of the research	236
7.4 – Political and pedagogical implications	237
7.5 – Suggestions for further research	238
REFERENCES	239
APPENDICES	247

LIST OF TABLE

TABLE	Page
1 – Summary of contrasts between past and current conceptions of validation	16
2 – Facet of the validity table by Messick (1989)	17
3 – Progressive matrix represented by Cummings (1996) and by Chapelle (1999).	18
4 – Areas of language knowledge, as presented in Bachman and Palmer (1996).	47
5 – Validity table presented by Chapelle (1994)	82
6 – Validity table with the extra column of the refutation of the argument against	83
7 – Specifications for university studies: general conditions for academic purposes based on Weir et al (2000)	96
8 – Specification for university studies – skills and strategies for each type of reading (Urquhart and Weir, 1998)	97
9 – Specification for university studies: purposes for skills and strategies (Urquhart and Weir, 1998)	97
10 – Definition of the tasks to be used for the criterion-related evidence	100
11 – Analysis of justifications for method MCQ as used within IELTS	117
12 – Analysis of the technical quality of item 5, test 1, IELTS	118
13 – Analysis of justifications for method short-answer question as used within IELTS	119
14 – Analysis of justifications for method sentence completion as used within IELTS	120
15 – Analysis of justifications for method completion of tables as used within IELTS	121
16 – Analysis of justifications for method completion of summaries as used within IELTS	122

17 – Analysis of justifications for method choosing headings as used within IELTS	123
18 – Analysis of technical quality of item 3, test 1, IELTS	124
19 – Analysis of justifications for method identifying writer’s claims, views or attitudes as used within IELTS	125
20 – Analysis of justifications for method classification as used within IELTS	127
21 – Analysis of justifications for method matching phrases/lists as used within IELTS	128
22 – Analysis of justifications for item 1 within TOEFL	131
23 – Analysis of justifications for item 2 within TOEFL	133
24 – Analysis of justifications for item 3 within TOEFL	134
25 – Analysis of justifications for item 4 within TOEFL	135
26 – Analysis of justifications for item 5 within TOEFL	137
27 – Analysis of justifications for item 6 within TOEFL	138
28 – Analysis of justifications for item 7 within TOEFL	139
29 – Analysis of justifications for item 8 within TOEFL	141
30 – Analysis of technical quality for item 9 within TOEFL	143
31 – Analysis of technical quality for item 10 within TOEFL	144
32 – Analysis of justifications for item 11 within TOEFL	145
33 – Analysis of justifications for item 12 within TOEFL	148
34 – Analysis of justifications for item 13 within TOEFL	150
35 – Analysis of justifications for item 14 within TOEFL	151
36 – Analysis of justifications for item 15 within TOEFL	153
37 – Analysis of justifications for item 1, EAP teacher 1	155
38 – Analysis of justifications for item 2, EAP teacher 1	156

39 – Analysis of justifications for item 3, EAP teacher 1	157
40 – Analysis of justifications for item 4, EAP teacher 1	158
41 – Analysis of justifications for item 5, EAP teacher 1	159
42 – Analysis of justifications for item 6, EAP teacher 1	160
43 – Analysis of justifications for item 2, EAP teacher 2	162
44 – Analysis of justifications for item 4, EAP teacher 2	163
45 – Analysis of justifications for item 6, EAP teacher 2	164
46 – Analysis of technical quality for item 8, EAP teacher 2	166
47 – Analysis of justifications for item 9, EAP teacher 2	167
48 – Analysis of justifications for item 11, EAP teacher 2	168
49 – Analysis of justifications for item 12, EAP teacher 2	169
50 – Analysis of justifications for item 13, EAP teacher 2	170
51 – Analysis of justifications for item 14, EAP teacher 2	171
52 – Analysis of justifications for item 15, EAP teacher 2	171
53 – Analysis of justifications for item 1, UFSC entrance examination	179
54 – Analysis of justifications for item 2, UFSC entrance examination	181
55 – Analysis of justifications for item 3, UFSC entrance examination	182
56 – Analysis of justifications for item 4, UFSC entrance examination	184
57 – Analysis of justifications for item 5, UFSC entrance examination	185
58 – Analysis of justifications for item 6, UFSC entrance examination	186
59 – Analysis of justifications for item 7, UFSC entrance examination	187
60 – Analysis of justifications for item 8, UFSC entrance examination	188
61 – Analysis of justifications for item 9, UFSC entrance examination	189
62 – Analysis of justifications for item 10, UFSC entrance examination	190
63 – Analysis of justifications for item 11, UFSC entrance examination ...	191

64 – Analysis of justifications for item 12 UFSC entrance examination	193
65 – Analysis of justifications for item 1, UNICAMP entrance examination	201
66 – Analysis of justifications for item 2, UNICAMP entrance examination	203
67 – Analysis of justifications for item 3, UNICAMP entrance examination	204
68 – Analysis of justifications for item 4, UNICAMP entrance examination	206
69 – Analysis of justifications for item 5, UNICAMP entrance examination	207
70 – Analysis of justifications for item 6, UNICAMP entrance examination	208
71 – Analysis of justifications for item 7, UNICAMP entrance examination	209
72 – Analysis of justifications for item 8, UNICAMP entrance examination	211
73 – Analysis of justifications for item 9, UNICAMP entrance examination	212
74 – Analysis of technical quality for item 9, UNICAMP entrance examination	212
75 – Analysis of justifications for item 10, UNICAMP entrance examination	213
76 – Analysis of justifications for item 11, UNICAMP entrance examination	215
77 – Analysis of justifications for item 12, UNICAMP entrance examination	216
78 – Analysis of justifications for item 10, 1999 UNICAMP entrance examination	218
79 – Analysis of technical quality for item 10, 1999 UNICAMP entrance examination	219
80 – Analysis of justifications for item 9, 2000 UNICAMP entrance examination	220
81 – Analysis of justifications for item 16, 2001 UNICAMP entrance examination	221

LIST OF APPENDICES

APPENDICES	Page
1 – IELTS Examination – test 1	247
2 – IELTS Examination – test 2	258
3 – IELTS Examination – test 4	266
4 – TOEFL Test	270
5 – TOEFL item 13	275
6 – TOEFL item 14	276
7 – TOEFL item 15	277
8 – Test by EAP teacher 1	278
9 – Test by EAP teacher 2	280
10 – Analysis of the correspondence of test tasks in relation to the criterion tasks of UFSC entrance examination	282
11 – UFSC Entrance Examination	286
12 – Analysis of the correspondence of test tasks in relation to the criterion tasks of UNICAMP entrance examination	294
13 – Item 9, UNICAMP 2000 Entrance Examination	299
14 – Item 2, UNICAMP 2000 Entrance Examination	300
15 – Item 16, UNICAMP 2001 Entrance Examination	301
16 – Item 14, UNICAMP 2002 Entrance Examination	302
17 – UNICAMP 1998 Examination	303
18 – Item 10, UNICAMP 1999 Entrance Examination	308
19 – Taxonomies of reading skills	309
20 – UFSC Specifications.....	308
21 – UNICAP Specifications	309

CHAPTER I

1.1 – Introduction

Testing has been a common practice in our society, and has been a decisive instrument for providing relevant evidence for decision-making, for mainly three purposes: selection, certification, and educational. For both selection and certification purposes, it functions independently of the educational process, providing information about the test takers' skills or abilities based on their performance on tests.

For educational purposes, in the classroom, testing may play an important role in the decision-making process for pedagogical choices. It may provide the teacher with information about learners' achievement relevant for decisions concerning what to teach again or next, or about how to develop future courses. Testing, in this case, may, thus, provide essential information for the process of evaluation¹.

Language testing (LT) has been part of this practice, mostly reflecting the same procedures underlying test construction and use generally, and has been extensively debated, within the context of second language teaching (Bachman, 1990; Bachman & Palmer, 1996; Bachman, Davidson, & Milanovic, 1996; Hughes, 1989, 2003; Ommagio Hadley, 1993; Allison, 1999; Bachman & Cohen, 1998; Weir, 1993; Brindley, 2001; Genesee, 2001; Brown, 2000), and, more specifically, for the assessment of reading ability (Bernhardt, 1991; Urquhart & Weir, 1998; Aebersold & Field, 1997; Weir, Huizhong & Yan, 2000; Alderson, 1990a, 1990b, 1996, 2000; Alderson, Clapham & Wall, 1995).

¹ Genesee (2001) defines evaluation as a process going beyond assessment, "to consider all aspects of teaching and learning" (p. 145), including considerations on how educational decisions can be informed by the results of assessment, and presents four basic components of evaluation, depicted in a cyclical relationship, where one influences the next: 1) purpose for the evaluation is articulated; 2) relevant information is identified and collected; 3) information is analyzed and interpreted; and 4) decisions are made, providing new information for the cycle.

Scholars and researchers in testing, today, agree about important aspects to be considered in test development and use such as practicality, reliability, and validity. Scholars and researchers also agree that the most important aspect is consideration of validity, in particular, of construct² validity³, since it addresses whether the evidence collected based on performance on the test can be used for valid interpretation of the ability being assessed as dictated by the construct, i.e., if inferences about ability can be made based on performance elicited through the test items.

In the first definition I was introduced to, validity was a characteristic of the test and testing instruments. A test was considered valid if it measured what it was designed to measure. Even in more recent publications, some scholars consider validity a characteristic of the test. Hughes (1989) says that “a test is said to be valid if it measures accurately what it is intended to measure” (p. 22), repeated in Hughes (2003). Brown (1987) uses valid test and validity of a test when talking about a valid test of reading ability, and about his view that the validity of a test is the most complex criterion (p. 221). For some scholars, thus, validity refers to a characteristic of the test or of the testing instruments. These definitions of validity as a characteristic of the test led me to the following research questions: *What have been the procedures used for the assessment of reading in English as a foreign language?*, and *To what extent are they valid instruments for assessing reading competence in English as a foreign language?*

My focus was, thus, on the investigation of the validity of the instruments of the tests. However, current views of validity have been proposed, changing its definition.

² According to the Dictionary of Language Testing, published by Davies, Brown, Elder, Hill, Lumley, and McNamara (1999), construct can be defined as “an ability or set of abilities that will be reflected in test performance, and about which inferences can be made on the basis of test scores. A construct is generally defined in terms of a theory; in the case of language, a theory of language” (p. 31). Ebel and Frisbie (1991, as cited in Alderson et al, 1995) define construct as a theoretical conceptualization of human behavior, which do not allow for direct observation, and give the examples of intelligence, motivation, anxiety, and reading comprehension.

³ According to the Dictionary of Language Testing, published by Davies, Brown, Elder, Hill, Lumley, and McNamara (1999), construct validity of a test “is an indication of how representative it is of an underlying theory of language learning [and] construct validation involves an investigation of the qualities that a test measures, thus providing a basis for the rationale of a test” (p. 33). This is further discussed in chapter 1.

In the 1980s, Samuel Messick, an influential scholar in the field of testing, showed two aspects of validity that caused profound changes in testing. The first aspect was that validity was not a characteristic of the test. As Messick (1989) puts it “tests do not have reliabilities and validities, only test responses do” (p. 14). Validity, according to him, is related to the inferences and actions based on the test score (responses), and requires empirical evidence and theoretical rationales in the support of their adequacy and appropriateness.

Based on this definition, it is possible to ask: are the inferences and actions based on the test score adequate and appropriate? If there is enough empirical evidence and theoretical rationales to give support to the hypothesis that the inferences and actions are adequate and appropriate, it is possible to say that the inferences and actions (based on the test score) are valid to the extent of the evidence⁴.

Investigating adequacy and appropriateness requires defining a reference, or domain of reference, which may be, in the case of testing, the content of a course, or the criterion representing the performance in the real world. Ultimately, since the content of a course and/or the criterion in the real world are, or should be, based on some construct, investigating adequacy and appropriateness should be referenced to a construct.

The second aspect of validity shown by Messick was that tests should not be analyzed only in terms of their internal characteristics for the simple fact that the main purpose of a test is to be used, and its use has been, according to him, somewhat neglected in earlier investigations of validity in testing. Moreover, since tests are to be used, they have consequences, both intended and unintended, and these consequences should be an integral part of a framework to analyze validity in tests. Adding that tests have consequences caused a shift of paradigm in language testing.

⁴ To the extent of evidence expresses the idea by Messick (1989) that validity is a matter of degree, not an all or none matter (p. 13).

As to the validation process, McNamara (2000) explains its more recent definition by comparing the validation process in testing to the way a trial is conducted. In his words, “there are thus two stages, each involving the consideration of evidence. First, the police make an investigation, and on the evidence available to them reach the conclusion...This conclusion is itself then examined, using an independent procedure” (p. 47).

According to the author, the two stages of collecting evidence in the case of a trial can be compared to the process of validation in testing. The first stage of validation in testing concerns the gathering of evidence of performance through the test items, based on which a conclusion is reached, represented in test score. The second stage involves scrutinizing the test procedures leading up to the conclusion. These two stages should, in the author’s view, be an integral part of test validation. In addition to that, he claims, test validation, like trials, involves thinking about the reasoning of the test, its logic, and examining empirical evidence from test trials or administration.

Scrutinizing the test implies, based on this new definition of validation, securing sufficient justifications for test interpretation and use. As Chappelle (1994) stresses, validation requires providing justifications pertaining to appropriate test interpretation, justifications referring to empirical evidence and theoretical rationales (p. 161).

This comparison to a trial, with the need to provide justifications, is also present in influential articles in testing and language testing when using the word *case*. Messick (1989) defines validation as a continuous process of collecting further evidence, and as being essentially “a matter of making the most reasonable case” (p. 13) in support of the test. Bachman and Palmer (1996), in turn, define construct validation as a process involving building a case in support of a specific interpretation with evidence justifying the interpretation (p. 22). It is possible to see, in both authors, the idea of a comparison between test validation and the procedures of a trial, in the expressions ‘making a case’ and ‘building a case’.

Thus, considering the importance of validity and validation, investigating validity in language testing, more specifically validity in tests designed to measure reading ability is the objective of this study. The analysis is based on tests used for different purposes: to assess proficiency as measured by two standardized tests of TOEFL (Test Of English as a Foreign Language) and IELTS (International English Language Testing System); for selection as measured by two university entrance examinations used by two Brazilian universities – UNICAMP (Universidade de Campinas), and UFSC (Universidade Federal de Santa Catarina); and for achievement as measured by classroom tests used by two English for Academic Purposes (EAP) teachers at UFSC, Florianópolis, SC.

1.2 – Research Questions

Based on this more recent definition of validity as a characteristic of the interpretation and actions based on the test, as most scholars in testing accept today, and on the view that validation (construct validation) requires examining carefully the meaning of the performance obtained on the elicitation devices used in a test, in this research I will be carrying out item analysis, and, when pertinent, analysis of the test content.

Thus, I will be looking at the test items⁵ as elicitation devices, and the whole test for test content, to analyze the kind of evidence that is possible to be collected through them, to see if they allow evidence appropriate for a valid interpretation of language ability (construct validity). That is, I will be scrutinizing the test procedures leading up to the test score, which, in turn, is used for interpretation and action. Ultimately, I will

⁵ I will be using ‘item’ rather than method or techniques because ‘item’ reflects the level of detail pursued in this research. That is, I want to go beyond analyzing the method used to analyzing each item of each method in terms of their defensibility as to their technical quality and quality of the evidence collected for validation inquiry.

be talking about defensibility of test items, where a defensible item can be argued to provide the evidence needed for adequate and appropriate inference and action, i.e., valid interpretation of ability and valid action based on the interpretation.

My general research question, now, is: to what extent do the test items allow for collecting evidence leading to a valid interpretation of the reading ability of the test takers and/or valid action based on the interpretation?. In the pursuit of carrying out this research, three research questions are posed:

- 1 - Are the items used within the TOEFL and within the IELTS defensible for their purpose of providing evidence based on performance for interpretation of reading ability in English as a foreign language?
- 2 - Are the items used by EAP teachers defensible for their purpose of providing evidence based on performance for interpretation of reading ability in English as a foreign language?
- 3 - Are the items used by the institutes in charge of the university entrance examinations in Brazil defensible for their purpose of providing evidence based on performance for interpretation of reading ability in English as a foreign language and for the action of selecting candidates for university studies?

1.3 – Motivation for the Study

Reading in a foreign language is a skill required of graduate students in Brazil. Assessing the reading competence of these students in Brazilian universities has, thus, become necessary, and this has been done through the so-called proficiency tests in a

foreign language, which have become a mandatory condition since 1969 at Brazilian universities (Wielewicki, 1997).

At UFSC, the proficiency tests have been regulated by its Resolução 10/CUn/97, of July 29th, 1997, of the Regimento da Pós-Graduação *Stricto Sensu*, Artigo 18, which says that all graduate students must take the proficiency test in a foreign language, without any further specification.

The method used for the proficiency tests at UFSC has been the translation of a short text, usually 30 lines. The use of translation as the method for assessing reading ability/competence in a foreign language seemed to be inadequate. Translating involves cognitive skills with specific task demands on the cognitive system, and reading, in turn, involves different cognitive skills with different demands.

I had suspected that the choice for translation as the method used to assess proficiency was made mostly considering the paradox involved in the nature of this kind of testing situation: since it has been shown by schema theorists that reading comprehension involves the reader's background knowledge in the development of a mental representation of the text, and since the test users and test takers are from different areas of expertise, the former from the Letras Department, and the latter from the various areas of the university, the outcome of the reading process might be argued to be somewhat different, impeding the use of methods or items focusing on comprehension.

The coordinator of the department responsible for the proficiency test at UFSC confirmed my suspicion, and added that comprehension questions had already been used previously with frequent events revealing the paradox, with the different people involved arguing in favor of their comprehension, events which caused them to stop the use of comprehension items, and return to the use of translation (personal communication, September, 2000).

The pursuit to understand the paradoxical situation aforementioned and how various tests deal with it was the starting point for this research.

1.4 – Significance of the Study

In general, the significance of this kind of research is to provide test users with information for the development or choice of the most adequate test or items for the construct to be assessed, especially because in validation theory today, it is the test users' responsibility to justify validity for any specific use of a test (Chapelle, 1999).

A theoretical significance is to provide a review of literature of the most recent research in the area of testing, in particular the new definitions put forth by Messick (1989), which has caused a paradigm shift in the area. Also, still in the area of testing as a whole, since this research can be an example of a study of validation based not on correlational studies as done in the past, but based on the argument-based approach with the focus on the justifications based on the constructs of language ability and reading ability used in this research.

Specifically for the area of language testing, its significance is justified by the use of the framework proposed by Bachman (1990) and Bachman and Palmer (1996) to investigate the authenticity of test tasks in comparison to the target language use tasks, which has been considered a very useful tool for test analysis and for any inference concerning future performance in the domain of reference.

The social significance of this research is that it shows the importance of pursuing validation in testing, in particular, because tests have consequences, both to the individuals and to society, both intended and unintended, as well as power, in particular because all the choices involved in the test development (included in their specifications) are determined by the test developers reflecting their values.

The ethical significance of the research is to show that tests may be unfair (biased), and to point to ways of developing tests which include features of fairness, features required in most codes of testing practice currently available.

The pedagogical significance may be more restrict, providing the teacher with information as to pursue validity in the development of their tests, and may be broader, providing the teacher with a discussion on the theoretical, social, and ethical issues involved in the development and use of a simple test.

1.5 – Organization of the Dissertation

The dissertation is organized in seven chapters. Chapter I includes an introduction to the problem investigated, the presentation of the research questions, the motivation for the study, the significance of the study, and the organization of the dissertation.

Chapter II brings a review of the literature related to validity in testing. I present a historical perspective of validity, with the evolution of the concept of validity, and the concept of validity in language testing, including its reinterpretation by Bachman and Palmer (1996), and I also introduce the framework by Bachman (1990) to determine the authenticity required for considerations of criterion validity. I end chapter II with a discussion, with an extreme perspective, of the impact tests have on society as whole and on individuals, therefore considerations of consequential validity.

In chapter III, I present the debate on the construct of language ability and of reading ability to be used in this research for the investigation of construct validity and validation, and the definition of the scope of the construct, including a discussion on what traditional testing may incorporate. I end the chapter with the debate on the definition of reading as a unitary skill or multidivisible skill, and the related discussion

on the possibility of establishing a relationship of correspondence between item used and skill assessed.

In chapter IV, I describe the method of the research, which includes a discussion on validation procedures proposed by scholars in the area. I also present a discussion on the sources of validity evidence, the methods for collecting validity evidence, and some definitions for this research: definition of the criterion, and definition of the characteristics of the framework for authenticity analysis. I, then, present the notion of Reverse Engineering proposed by Davidson and Lynch (2002) as a confirmation of the approach adopted in terms of the direction of the analysis, from the test items to the specifications to the construct assessed. The material collected for analysis is offered in the last section.

In chapter V, I present the analysis, with evidential basis, of the data collected from the TOEFL test and the IELTS test, and from the EAP teachers. In chapter VI, I present the analysis, with evidential basis and consequential basis, of the data collected from the university entrance examinations - UNICAMP and UFSC.

And finally, in chapter VII, the main conclusions of the study, the final remarks, the limitations, some political and pedagogical implications, and suggestions for further research, are presented.

CHAPTER II

Conceptions of Validity in Testing

In this chapter, I present the review of literature in testing. The first section (2.1) presents a review of the most important studies on validity in testing, the second section (2.2), the debate on validity for language testing, and the third section (2.3), a discussion on consequential validity with an extreme perspective.

2.1 – Validity in testing

In this section, I will show the evolution of the concept of validity from a historical perspective (2.1.1), and the more recent definition of validity as a unitary concept (2.1.2).

2.1.1 – Historical Perspective of Validity: the evolution of the concept.

Validity has been subject of debate and discussion, definition and redefinition for many years. Cumming (1996) mentions 16 types of validity since the 1930s: concurrent, construct, content, convergent, criterion-related, discriminant, ecological, face, factorial, intrinsic, operational, population, predictive, task, temporal, and validity generalization.

Validity dates back to the time philosophers were concerned with validation within scientific method. Messick (1989) shows that philosophers such as Leibnezi, Lock, Kant, Hegel, and Singer presented their different modes of inquiry for the

scientific method with considerations of the process of collecting validity evidence, that is, the process of validation within the scientific method.

In the area of testing, validity appeared first as early as the 1920s, when testers were already looking for evidence of reliability and validity in intelligence tests in the USA. A famous tester at the time, named Carl Brigham, was already concerned about the validity of tests used to measure intelligence at the time (Lemann, 1999).

From the 1920s to the 1950s, validity was defined as anything with which a test could correlate (Shepard, 1993). In the 1940s, the usual definition of validity at the time was still: “a test is valid for anything with which it correlates” (Guilford, 1946, as cited in Messick, 1989, p. 18).

By 1949, however, it was already possible to notice changes in the definition of validity and validation with the publication of Cronbach’s 1949 book, with the idea of plausible rival hypothesis in his notion of logical validity (Messick, 1989), under the influence of the philosopher Popper and his idea of falsifiability in the area of testing. Construct validity was introduced in 1954 (Shepard, 1993) as a result of the effort of the American Psychological Association while preparing the code of professional ethics. In 1955, a conceptual framework for its investigation was already put forth in the article by Cronbach and Meehl (1955, as cited in Bachman, 1990). In 1957, plausible rival hypothesis as a validation procedure was mentioned in Campbell (1957, as cited in Messick, 1989, and in Shepard, 1993).

Until the early 1960s, validity was considered to be of four types: content, predictive, concurrent, and construct. Messick (1989) provides the definitions of the four types of validity at the time. Content validity refers to conclusions based on the evaluation of a comparison between test content and the situations or subject matter of interest. Predictive validity refers to how well the test can predict performance in the

situation/criterion of interest. Concurrent validity refers to how well the test estimates the present standing of the individual on the criterion of interest. And finally, construct validity refers to an investigation of the qualities measured by a test which are indicative of the construct.

However, with the publication of the American Psychological Association (APA) Standards in 1966, predictive and concurrent became one: criterion-related validity (Shepard, 1993), defined, at the time, as evaluation by comparison of test score with some external variable reflecting the characteristics of the behavior of interest (Messick, 1989). Thus, through the 1960s and 1970s, validity, construct, content, and criterion-related were considered the Holy Trinity, because they assumed the character of religious orthodoxy (Shepard, 1993).

In the 1974 version of the APA standards, the interrelatedness of the three types of validity was recognized. A unified conception of validity, under the construct validity framework, was being proposed by scholars such as Lee Cronbach, Anne Anastasi, and Samuel Messick.

In 1980, Samuel Messick contributed with the idea of considering social values and consequences into the concept of validity, and also with the idea that the “meaning of the measure, and hence its construct validity, must always be pursued – not only to support test interpretation, but also to justify its use” (Messick, 1989, p. 17).

Still in the 1980s, there was an emphasis by the three scholars that the three types of validation should not to be taken as alternatives to one another, but rather, should be taken all together to provide an integrated explanation, and, since the goal of validation is explanation and understanding, “all validation is construct validation” (Cronbach, 1984, as cited in Messick, 1989, p. 19). Thus, in the early 1980s, it is already possible

to see an increasing emphasis on construct validity as the essence of a unitary validity conception.

In the 1985 APA standards, validity was defined as “the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores” (American Psychology Association, 1985, as cited in Bachman, 1990, p. 243), and it was already seen as a single unified concept, with construct validity as the central concept.

In 1988, Lee Cronback published his influential article titled *Five Perspectives on Validity Argument* in which validation was considered as persuasive argument, according to which validation should include a debate of pros and cons arguments to defend the interpretations which can be drawn from a test. It is possible, then, to see in his article, a clear movement towards validation as a persuasive argument.

In 1989, Samuel Messick published a seminal paper with the title *Validity*, in which he defines validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment” (p. 13).

Messick’s extensive treatise, as it is known to most researchers in testing, a) cements the consensus that construct validity is the one unifying conception of validity, and b) extends the boundaries of validity beyond test score meaning to include relevance and utility, value implications, and social consequences (Shepard, 1993).

In sum, as Messick (1989) points out, there are three moments in the definition of the concept of validity. The first, when validity was defined very vaguely as anything with which a test can correlate. A second moment, when four types of validity were conceived – content-, concurrent-, predictive- and construct- validity, which later were reduced to three: concurrent and predictive subsumed into criterion validity, content

validity, and construct validity. The third moment, which has the most influence on testing today, when researchers increasingly consider validity as one.

The evolution of the concept of validity took place over some years, with scholars such as Anastasi, Cronbach, and Messick contributing decisively. The idea that Messick (1989) “cements the consensus” as put forth by Shepard (1993) is important since, in his article, the author, in fact, brings together the main ideas from the debate around validity by many scholars for decades, thus encompassing the evolution of the definition of validity to that date.

The main contribution by Samuel Messick in the 1980s was to advocate social consequences of test interpretation and test use as an integral part of the validity argument, thus reminding the world that tests are for use with people (Shohamy, 2001), a dimension which would require further analysis of the values, ideologies, and broader theories in relation to the conceptual framework used (Cumming, 1996).

Chapelle (1999) presents a table where it is possible to compare the past conceptions of validation to the current ones, i.e., before and after the 1980s. It is presented below as table 1.

Table 1: Summary of contrasts between past and current conceptions of validation

Past	Current
Validity was considered a <i>characteristic of a test</i> : the extent to which a test measures what it is supposed to measure.	Validity is considered an <i>argument</i> concerning test interpretation and use: the extent to which test interpretations and uses can be justified.
Reliability was seen as distinct from and a necessary <i>condition for validity</i> .	Reliability can be seen as <i>one type of validity evidence</i> .
Validity was often established through <i>correlations</i> of a test with other tests.	Validity is argued on the basis of a number of types of <i>rationales and evidence</i> , including the consequences of the testing.
Construct validity was seen as one of <i>three types of validity</i> (the three validities were content, criterion-related, and construct).	Validity is a <i>unitary concept</i> with construct validity as central (content and criterion-related evidence can be used as evidence about construct validity).
Establishing validity was considered within the purview of <i>testing researchers</i> responsible for developing large-scale, high-stakes tests.	Justifying the validity of test use is the responsibility of <i>all test users</i> .

2.1.2 – Validity as a unitary concept: the progressive matrix proposed by Messick

Messick (1989) solidifies the previous changes by providing a matrix of facets of validity for the argument-based validation and by proposing what has come to be known as the unified validity framework. This framework was developed by using two distinct though interconnected facets of the unitary validity concept: one is the source of justification of the testing, based on either the evidence or the consequence of the testing procedures; and the other is the function or outcome of the testing, involving either the interpretation or use of the test score.

Distinguishing the facets would result in the four-fold classification Messick has proposed for the testing field, which is presented in table 2 below.

Table 2: Facets of the validity by Messick (1989)

	Test interpretation	Test use
Evidential basis	(1) Construct validity	Construct validity + (3) Relevance/utility
Consequential basis	(2) Value implications	(4) Social consequences

The author stresses that it is a progressive-matrix formulation, since the next cell includes the content of the previous. Thus, in the first column, for test interpretation with evidential basis, there is construct validity (1)⁶, and with consequential basis, there are construct validity (1) + value implications (2). In the second column, for test use with evidential basis, there are construct validity (1) + relevance/utility (3), and with consequential basis, there are all of them: construct validity (1) + relevance or utility (3) + value implications (2) + social consequences (4). It is possible to see that construct validity will appear in every cell of the table, “thereby highlighting its pervasive and overarching nature...taken as the whole of validity in the final analysis” (Messick, 1989, p. 21).

To better illustrate the progressive matrix of construct validity or validation with construct validity as part of every cell, a modified table introduced by Cumming (1996) and by Chapelle (1999) is presented in table 3 below.

⁶ These numbers were introduced in the cells by this researcher for the purpose of explanation, thus not belonging in the original table.

Table 3: Messick’s (1989) progressive matrix of construct validation represented by Cummings (1996) and by Chapelle (1999).

Source of justification	Function of outcome of testing	
	Test interpretation	Test use
Evidential basis	Construct validity	Construct validity + Relevance/utility
Consequential basis	Construct validity + Value implications	Construct validity + Relevance/Utility + Value Implications + Social consequences

Messick’s formulation shows that one facet is the source of justification, based on “appraisal of either evidence or consequence” (Messick, 1989, p. 20). It is possible to see, based on the first column, that two sources of justification must be provided for test interpretation: evidence for the construct validity must be gathered and the value implications as to what to implicate in the construct, and why, must be considered.

It can be seen, based on the second column, that two sources of justification must be provided for test use: evidence for the construct validity and for relevance or utility for a specific use must be gathered. Also, the value implications and the social consequences of using a particular test within that construct and with evidence of relevance for a specific use must be considered.

All this distinction is not, however, so clear-cut. Messick himself recognizes the difficulty of distinguishing the sources of evidence the way explained in his matrix, and also of making use of them for interpretation and use, difficulty explained by the fact it is a unitary concept. As the author himself says, “the fuzziness – or rather the messiness – of these distinctions derives from the fact that we are trying to cut through what

indeed is a unitary concept” (p. 21). The author stresses, however, that the distinction is important, since it makes it easier to appraise the appropriateness, meaningfulness, and usefulness of the interpretation drawn based on the test score.

2.2 – Validity in Language Testing

In 1961, Robert Lado published a book, very influential at the time, called *Language Testing*, where validity is defined as “does the test measure what it claims to measure? If it does, it is valid” (p. 321). In the book, validity was reported mainly by correlation with some valid criterion, which could be another test, or any other assessment procedure considered valid by an expert or competent judge. Validity, for him, was a characteristic of the test, an all-or-nothing attribute (Chapelle, 1999).

In the 1970s, validity in language testing was being redefined, since the trend for communicative competence led researchers to “probe questions about the construct validation for tests of communicative competence” (Chapelle, 1999, p. 255). In the 1980s, although the discussion in the field of education was very explicit in terms of the definition of validity, the discussion in language testing was not very explicit for the definition and scope of validity (Chapelle, 1999). There were some researchers with the more traditional definitions of validity, such as Hughes (1989), and Brown (1987), as comprising of four different types of validity: content, construct, predictive, and face validity.

However, there was already a movement towards the redefinition of validity. Chapelle (1999) points out that, in the 1980s, some researchers were investigating processes for validation other than the correlational method by trying to understand the method of hypothesis-testing while others were suggesting the inclusion of new ideas

as part of the concept of validity, such as the idea of affect, of ethics, of washback, and the idea that a test is valid for a specific purpose.

Although it is possible to see the notion of test consequences already mentioned by scholars as concern about affect, washback, and ethics as pointed out by Chapelle (1999), it was Bachman (1990), through his chapter on validity, who had the most influence in language testing, in particular by defining validation as a process of presenting a variety of evidence about test interpretation and use (Chapelle, 1999).

2.2.1 – Bachman (1990) interpreting validity to language testing

Bachman (1990) took on the task of interpreting the new definitions of validity and validation put forth by Cronbach and Messick in the 1980s into language testing. The author accepts: a) that use must be considered for validation, and b) that validation requires all relevant information. He accepts that the use of the test must be considered in the process of validation, and with it, the value systems that justify the use of the test. Thus, ethical values should be investigated in collecting validity evidence.

The author also accepts that, in validation, it is not the test content or the test score which are being examined, but the way the information gathered through the testing procedures can be interpreted or used. This requires, according to the author, reference to the specific ability or abilities the test is designed to measure, which, in turn, requires a theory of abilities for the identification of sources of variance, and the uses for which the test is intended. Thus, in investigating validity, factors making up language ability and factors other than the language abilities affecting performance must be examined.

Also for validation, the performance must be measured in a reliable way, without errors in measurement. As Bachman (1990) puts it, “if we demonstrate that test scores are reliable, we know that the performance on the test is affected primarily by factors other than measurement errors” (p. 236). Reliability must, thus, be seen as a requirement for validity.

The resulting statement of the validation process should answer the question to what extent the test analyzed provides a valid basis for making inferences about language ability. This process should take into consideration not only the individuals, but also the context involved in the use of language, including the collection of evidence as support of a given interpretation or use, which is based on logical, empirical, and ethical considerations (Bachman, 1990). Moreover, this process should involve raising hypotheses of evidence, but also counterhypotheses.

Based on this, it is possible to understand the new role of validation in testing since the early 1980s, which is accepted by Bachman (1990) for language testing with the following quotation: “the job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only when it has survived serious attempts to falsify it” (Cronbach, 1980, as cited in Bachman, 1990, p. 257). This new role of validation has determined the method used in this research (as explained further in chapter 4).

Bachman looks at all the aspects put forth by Messick in his progressive matrix, and divides the sources of evidence for validation into evidential basis and consequential basis of validity.

Evidential Basis of Validity

According to Bachman, the evidence collected, to support the decision for a particular use of a test, can be grouped in three types: a) content relevance and content coverage; b) criterion relatedness; and c) meaningfulness of construct.

Both content relevance and content coverage, the first type of evidence, must be demonstrated in the process of validation, which requires that the specifications of the test in terms of domain of abilities and underlying tasks be clearly and unambiguously identified.

Content relevance refers to two aspects: 1) the relevance of the domain, the language ability or abilities in the case of language tests; and 2) the methods – items or tasks – used to elicit the performance. Content coverage is related to the extent to which the tasks required in the test adequately represent the behavioral domain in question. Although a necessary part of the validation process, examining content relevance and content coverage is not enough for inferences of language ability.

Criterion refers to a domain of behavior, for example, a framework of language ability or some proficiency scale to be used as a reference against which the performance on the test is compared. Thus, criterion relatedness, the second type of evidence, requires identifying an appropriate criterion behavior and demonstrating that scores on the test are functionally related to this criterion.

The validation process can demonstrate a relationship between test score and some criterion by collecting information about two testing situations occurring almost

at the same time, concurrent validity⁷, or information believed to be needed for future use, predictive validity.

Predictive utility/validity is related to how well the test can predict some future behavior, thus concerned with the accuracy with which test scores predict the criterion behavior of interest, which requires, for example, demonstrating a relationship between test scores and performance in the criterion. Although a necessary part of the validation process, examining the correspondence between performance on the test and performance in the criterion is not enough for inferences of language ability, that is, valid predictors of future performance are not valid indicators of ability.

Construct validation, i.e., the meaningfulness of construct, the third type of evidence, is the most fundamental of the three types, since construct validation must be demonstrated if the results of the tests are to be interpreted as indicators of ability. Construct can be defined as the abilities identified which “permit us to state specific hypotheses about how these abilities are or are not related to other abilities, and about the relationship between these abilities and observed behavior” (Bachman, 1990, p. 255). The importance of defining the construct relies on the fact that the construct is used for all interpretation of abilities and, thus, for construct validation.

Construct validity can, thus, be defined as the extent to which inferences about hypothesized abilities based on test performance can be made, and construct validation can be compared to a “special case of verifying, or falsifying a scientific theory, and just as a theory can never be ‘proven’, the validity of any given test use or interpretation is always subject to falsification” (Bachman, 1990, p. 256).

⁷ Concurrent validity is a kind of criterion-related investigations which takes place almost at the same time as the test itself, and can take two forms: a) examining differences in test performance among groups of individuals at different levels; b) examining correlations among various measures of the same ability, usually correlating with some standardized tests. It will not be worked with in this research.

Construct validation, according to author, requires both logical analysis and empirical investigation. Logical analysis is involved in defining the construct theoretically and operationally, implying that hypotheses are made at both the theoretical and operational levels.

As to empirical investigation, Bachman presents three types of empirical evidence to support construct validity: correlational (quantitative), experimental (quantitative), and process analysis (qualitative). Since process analysis allows the investigation of the processes involved in the performance of tasks, hence allowing for a better understanding of what the test takers actually do when taking the tests, and, as a consequence, what the language tests actually measure, it will be part of the method for collecting evidence of construct validity in this research, and is further explained in chapter 4.

Another aspect Bachman considers important for evidential basis of validity is test bias. According to the author, any validation processes should involve considerations as to test bias when collecting evidence, since it may affect the way test takers perform on the test, and may result in inaccurate information. Differences in performance should be related to differences in language ability, providing valid evidence for the interpretation of ability, not related to test bias.

The following four aspects are presented as potential sources of bias, thus as potential sources of invalidity when not considered or not included in the construct: a) cultural background; b) native language, ethnicity, sex, age; c) background knowledge; and d) cognitive characteristics (field independence or ambiguity tolerance).

Although all these sources of invalidity are relevant for any test development, since they all might, one way or another, affect performance on language tests, for the

present research, background knowledge is the only source relevant for discussion. Background knowledge, for Bachman, refers to prior knowledge of the content in tests, and may be a source of bias when favoring groups or individuals with more familiarity with the topic. The author stresses that, although it is difficult to distinguish between background knowledge and language ability, this distinction must be made, since it has “crucial implications for the development and use of language tests and for the interpretation of their results” (p. 274).

Bachman stresses that while cognitive characteristics and characteristics such as native language, age, sex, ethnicity cannot be considered as part of the construct being measured, background knowledge can. In the former case, they cannot be considered as facets of the test methods since they cannot be taken out. In the latter case, the test must be developed according to the construct defined, and the differences in the test results with individuals with different backgrounds must be considered “not as instances of test bias, but as indicators of different levels on this specific ability to be measured” (p. 279).

In defining the scope of the construct, decisions as to what factors to include and also what factors to not include must be made. As the author points out, the interpretation of the effects of these factors as “source of measurement error, test bias, or part of the language abilities we want to measure will depend on how we define the abilities and what use is to be made of test scores in any given testing situation” (p. 278).

All the factors aforementioned may affect performance. Thus, defining the scope of the construct, i.e., not only what to include, but also what not to include, is one of the most important aspects of a test development, since it dictates what can be considered

valid or invalid, and affects how the interpretations based on the scores can be made. This is further discussed in section 3.3.

Consequential or ethical basis of validity (consequential evidence)

Tests have been developed and used to serve the needs of an educational system or of a society as a whole, and their use should be appropriate in terms of the intended consequences, and responsible in terms of the potential unintended consequences.

In considering the validity of test score and test uses, test developers and test users must, according to Bachman (1990), consider all the ethical and political issues, going “beyond the scientific demonstration of empirical and logical evidence to the consideration of the potential consequences of testing...mov[ing] out of the comfortable confines of applied linguistic and psychometric theory and into the arena of the public policy” (p. 281).

The author addresses the four areas considered by Messick (1989) in the ethical interpretation and use of test results to determine the appropriate use of a test: construct validity, value system, practical usefulness of the test, and the consequences to the educational system and society.

In agreement with Messick, Bachman (1990) claims that construct validity must be considered for the evidence supporting a particular interpretation, and that value system must be considered since test developers and users all have value systems used to make decisions such as the balance between demands of reliability, validity, and practicality considering the resources, or the criterion to be used as reference.

Practical usefulness of the test must be considered so as to provide evidence to support the test use in terms of the relevance of the tasks which should be contemplated

within construct validation. Consequences of using test results for a particular purpose to the educational system and to society must also be considered to determine whether the benefits are as expected and outweigh the negative consequences for both individuals and society.

All the decisions must consider a balance between the political component determining the values to be implemented and the psychometric component determining how the values will be implemented. To justify a particular interpretation, evidence for construct validity must be gathered and the value implications of this interpretation must be considered. To justify the use of scores from a test as a basis for decisions, evidence must be presented or argued as to its relevance, and the consequences of the decision must be considered. Unlike a traditional question such as *what does the instrument measure?*, a judgmental question would be: *why should that be measured?*.

2.2.2 – Alderson et al (1995) interpreting validity to language testing

Alderson et al (1995) write a whole chapter on validation, in which they divide validity in three distinct types: internal validity, external validity, and construct validity. Internal comprises face validation, content validation, and response validation. External validity comprises concurrent validity and predictive validity. Construct validity is collected by using correlations, through comparing with some theory, with some internal component (such as scores both in reading and writing), with some other characteristics (such as test takers' biodata or psychological characteristics), and by using factor analysis (showing correlation of a smaller number of factors).

Based on the definitions of validity and validation Alderson et al (1995) present in their chapter, it is possible to come to several conclusions. The first is that they recognize the discussion which considers construct validity to be the important type of validity, the unifying element, to which the other types of validity should contribute by providing evidence. As they assert, some researchers in testing “believe that it [construct validity] is a superordinate form of validity to which internal and external validity contribute” (p. 183). The authors do not, however, clarify whether or not they accept that notion of construct validity as the central element.

The second conclusion is that they still use face validity, which is, according to them, neglected on the grounds that it is “unscientific and irrelevant” (p. 172). They state, however, their opinion on the issue: “face validity is important in testing”(p. 173), since the test taker must take the test seriously or else the response validity is affected.

The third conclusion is that they do not incorporate in their conception of validity an explicit concern for the consequences of the tests. Aside from mentioning that the test scores must be shown to be a “fair and accurate reflection of the candidate’s ability” (p. 188), where fair implies considerations of consequences, there is no other explicit mention of the consequences as put forth by Messick (1989) in educational psychology and Bachman (1990) in language testing, as explained above.

The fourth conclusion is that they still consider validity as investigated through correlations, internal components or other characteristics, being, thus, very distinct from the definition of validity and validations as put forth by Messick (1989) and Bachman (1990), as already mentioned.

2.2.3 – Bachman and Palmer’s (1996) notion of usefulness and framework of language task characteristics

Bachman and Palmer (1996) accept the definition of validation as making a case and providing justification for an interpretation. They define construct validation as the “on-going process of demonstrating that a particular interpretation of test scores is justified, and involves, essentially, building a logical case in support of a particular interpretation and providing evidence justifying that interpretation” (p. 22).

However, while recognizing the central role of construct validity, they do not endorse completely the framework proposed by Messick (1989), which is possible to see when they make clear that other qualities are “important enough to the development and use of language tests to warrant separate consideration...as separate qualities” (p. 42). The qualities mentioned by the authors are reliability, interactiveness, practicality, impact, construct validity, and authenticity. These qualities are described under the notion of usefulness, which is, according to them, the most important quality, since a test is to be used for something.

Reliability refers to consistency of measurement, i.e, if the test taker has the ability being measured, he or she must be able to demonstrate the ability in all testing situations. It is, thus, a “function of consistencies across different sets of test task characteristics” (p. 20). Interactiveness refers to the kind of involvement of areas such as language ability (language knowledge and strategic competence), topical knowledge, and affective schemata, being more or less interactive depending on the task requirements. Practicality refers to the resources required for the use of the test, and is considered relevant since it may affect the decisions concerning all the other qualities.

Impact, the fourth quality, may be on the individuals (test takers and teachers), and on the society and educational systems. Impact on test takers is expressed in terms

of fair decisions and fair test use, where the former refers to equal appropriate treatment to all test takers regardless of their group membership, and the latter to relevance and appropriateness of the score to the decision. Impact on the teachers is usually manifested in their having to develop their program, including a “teaching to the test” syllabus, with the negative aspect of being forced to give up their autonomy for class decisions. Impact on society is related to the values and goals influencing the decisions and actions based on the test score. Considerations of impact require reflecting upon the possible consequences, both positive and negative, of using a specific test, not an alternative, for the same purpose.

Construct validity, the fifth quality, refers to the meaningfulness and appropriateness of the test score interpretation. It requires adequate justifications for the interpretation based on the score, which involves defining a construct of language ability, and providing “evidence that the test score reflects the area(s) of language ability we want to measure, and very little else” (p. 21). In practical terms, this means that getting an item right should be evidence of ability, or abilities as defined in the construct. Conversely, getting the item wrong should be evidence of lack of the ability.

In addition to the construct, adequate justifications require determining the correspondence between the testing situation and the target language use (TLU) situation. The construct of language ability is discussed in chapter 3. The discussion of the correspondence between the testing situation and the target language use (TLU) situation is below.

Authenticity, the sixth quality, is defined as “the degree of correspondence of the characteristics of a given language test task to the features of a TLU task” (Bachman & Palmer, 1996, p. 23), and allows the investigation of the extent to which the interpretations based on the test performance can be generalized to situations other than

the test itself. High degree of correspondence between the test task characteristics and the TLU task should be expected in a test for any investigation of validation, in particular to construct validation, since, ultimately, it is the construct based on some framework of language ability which will give interpretation to performance on the test.

The authors claim that the qualities should be seen as unified, since each one will “contribute in unique but interrelated ways to the overall usefulness of a given test” (p. 18). According to them, three principles should be used for the operationalization of usefulness: a) the overall usefulness is to be maximized; b) the individual qualities cannot be evaluated independently, but combined; and c) the balance among the qualities is to be determined for each situation.

Based on these principles, it is possible to notice that, although Bachman and Palmer (1996) do not endorse the framework by Messick (1989), they do recognize that there is an interrelationship among the qualities, thus endorsing the idea that they are unified. In this case, usefulness is the unifying element, rather than validity or construct validity. This substitution is confirmed by Chapelle (1999) in saying that “they [Bachman and Palmer (1996)] substitute ‘usefulness’ for ‘validity of score-based inferences and uses’” (p. 264).

The definition of the usefulness in terms of reliability, construct validity, authenticity, interactiveness, impact, and practicality are all part of a process of quality control considered essential. The authors add, however, that this quality control is only one of the two fundamental principles of their approach for the development and use of language tests. The other fundamental principle is the need for a correspondence between the test performance and language use. Usefulness, in their viewpoint, must be argued in terms of the correspondence between the language test performance and language use, and this correspondence must be part of the validity argument.

Showing the correspondence between the language use situation and the test use situation is fundamental if inferences are to be made about some individual's language ability and decisions are to be made based on those inferences. The degree of this correspondence will determine the extent to which the task characteristics in the test mirror the task characteristics in the language use situation, and will "determine, to a large extent, the authenticity of the test task, the validity of inferences made, and the domain to which those inferences will generalize" (Bachman & Palmer, 1996, p. 43). Only a high degree of correspondence between the test tasks and the tasks in the language use situation will allow the test users to generalize to competence the information they have based on test takers' performance elicited by the test tasks.

Bachman and Palmer (1996) propose a framework as a description of the characteristics of language testing methods based on Bachman (1990) which, in turn, is based on previous research in the area, "an extension and recasting, to a large extent, of Carroll's and Clark's taxonomies" (Bachman, 1990, p. 117).

Considered the most recent thinking about the relationship between language use and test design (Alderson, 2000, p. 164), the framework will, ultimately, help in the understanding of the effects of task characteristics, both for language use tasks and for test tasks, thus, helping language test developers to investigate the degree of correspondence between the target language use tasks and the test tasks.

Since the framework is very comprehensive, its characteristics are briefly described below, and the characteristics to be used for the analysis in this research will be further explained in chapter 4, where the method of this research is expounded. They are characteristics of: a) the setting; b) the rubrics; c) the input; d) the expected responses; and e) the relationship between input and response.

The setting is related to the physical circumstances of language use, and its characteristics are: a) physical characteristics, such as location, noise level; b) participants, with their roles in the meaning negotiation; and c) time of task, considering fatigue and freshness.

The rubrics are related to the information given to the test takers as to what they are expected to do, how they are expected to proceed with the test and tasks. They are: a) structure - information concerning how the test is organized in terms of parts, sequences, relative importance, and number of tasks/items; b) instructions, which must be explicit and clear allowing best performance for inferences to be made; c) time allotment, which dictates whether the test will be speeded or power test; and d) scoring method, with the information criteria for correctness, procedures for scoring the responses, and information concerning how the test score will be presented.

The rubrics will provide the test taker with the precise information as to what they are expected to do, how they are expected to proceed with the test and tasks. In this case, instructions and specifications of procedures and tasks must be clear so that the test taker will be able to perform at his/her best, and show their level of competence. The test taker must be able to get the item either right or wrong because the item is effectively eliciting performance of the construct being assessed, not because the test taker could not understand what he/she was expected to do, or how to proceed the task.

Since familiarity with test items or methods may influence test takers' performance on the test, it is wise to provide more information or more examples in case there is the possibility of unfamiliarity with the item or method used. Also, the language used in the rubrics must be simple and, when possible and necessary, the test takers' native language should be used.

The criteria for correctness must also be clear as to provide the test taker with the information on how he/she is expected to respond, and how much information to provide in the answer for it to be considered correct by the test raters. The number of tasks must also be clear so that the test taker will find the most efficient way to be organized to perform at his/her best. Although important for the test, these characteristics under rubrics are recognized by the proponents to be the ones allowing the least correspondence between tasks required for language use and tasks required to complete the test since these characteristics are hardly present in language use situation.

The input is related to all the source material the language user will use for the accomplishment of the tasks, both in language use situation and in test use situation. The characteristics of the input are: a) format – how the input is presented in terms of channel, form, language, length, type of input, degree of speedness, and vehicle; and b) language of input – related to the nature of the language used in terms of language characteristics, both organizational and pragmatic, and in terms of topical characteristics, such as cultural, academic, technical information.

The input is the material the test taker will be reading to respond to. In terms of the reading tests, this involves the text as well as the task required by the item. Language use task, according to the authors, refers to activities to achieve a specific objective in a particular situation. Form of the input (language, non-language, or both) refers to what form(s) of language are to be used in the input. This might include the written text, or illustrations in general, such as pictures, diagrams, etc.

The type of input involves both items and prompt, the former requiring selected or limited production response, and the latter extended production response⁸. They are

⁸ Selected, limited, and extended production responses will be explained under characteristics of expected responses below.

different types of tasks designed with the intent of eliciting some sort of response on the part of the test takers.

The language characteristics of the input can be analyzed following organizational characteristics (grammatical and textual), and pragmatic characteristics (functional and sociolinguistic). They are included in this framework for the purpose of guiding the analysis of the correspondence between real life situation and testing situation. Topical characteristics refer to the kind of information in the input.

The expected response refers to response the test takers are supposed to provide based on the instructions, and on the type of input presented in the test. Characteristics of the expected responses are very similar to the characteristics of the input. They are: a) format – how the response is to be produced in terms of channel, form, language, length, type (selected, limited production, extended production), and degree of speedness; and b) language of the response – related to the nature of the language used in terms of language characteristics, both organizational and pragmatic, and in terms of topical characteristics, such as cultural, academic, technical information.

In both cases of language of input and language of the expected response, the organizational characteristics are grammatical (vocabulary, syntax, phonology, graphology), and textual (cohesion, rhetorical or conversational organization). Pragmatic characteristics are functional (ideational, manipulative, heuristic, imaginative), and sociolinguistic (dialect/variety, register, naturalness, cultural references and figurative language).

And finally, the relationship between input and response refer to how input and expected response are related to one another, and its characteristics are: a) reactivity – the extent to which the input or response affect subsequent input or responses; b) scope of relationship – the amount or range of input necessary for the response to be elicited

as expected; and c) directness of relationship – the extent to which the test taker will be able to provide the expected answer based primarily on the information given in the input.

In sum, taking all the six qualities proposed by Bachman and Palmer (1996) explained above, in this research I will focus on construct validity, authenticity, and impact. Concerning construct validity, it is possible to see that Bachman and Palmer (1996) retain a similar conception of validity and validation, as proposed by Messick and the other scholars, meaning ‘building a case’ in support of a specific interpretation with the evidence justifying the interpretation.

Authenticity, for the authors, replaces relevance/utility as used by Messick (1989) with the advantage of having a framework for the analysis of the degree of correspondence of the language use situations. Impact is related to fairness of the decisions and uses of the test, and is similar to Messick’s concern with consequential validity.

Considering consequential validity or impact is more related to a political agenda than a technical one, according to Willingham (1999). It entails giving up the uses tests have been put to over the years. This is further discussed in section 2.3 below, under consequential validity to the extreme.

2.3 - Consequential Validity to the Extreme

In an announcement previous to a feast for a celebration, a school principal says “...as a school treat, all exams have been cancelled” (film “Harry Potter and the Chamber of Secrets”). In response to that, the principal was applauded cheerfully and vividly by all the hundreds or thousands of the school’s students, showing not only

their approval for the decision, but also, and more importantly, their feelings towards exams. The aversion of the kids to the school exams is to be considered when investigating validity in testing, in particular its consequential validity.

Exams, or tests, do not have a very beautiful history. Quite the opposite, tests are related to power, punishment, fear, anxiety, failure, and possibly many other such sentiments which are remembered for many years. They have been used: a) as gatekeepers, to grant permission to enter or to exit, the legitimate tool for inclusion of some and exclusion of others; b) to raise educational level; c) to control knowledge and entrance (Shohamy, 2001).

The first large scale administration of a test in history was when, during the first World War, nearly 2 million recruits were classified using IQ tests in USA as a preparation effort for the war. Large-scale IQ tests have been associated with the Eugenics Movement occurring during the first half of the 1900s in the USA, which had as its main objective to use selective breeding technique to improve intelligence, and other attributes, of human race (Lemann, 1999). The idea to use the IQ test for university entrance examination was that higher education should be reserved for highly intelligent people, as indicated by their IQ scores and “society should be classified according to brainpower, and the brainiest people should be its leaders” (Lemann, 1999, p. 24).

Tests have had this role in our societies, in particular university entrance examinations. The consequences of the use of tests have been analyzed by Shohamy (2001), who has taken this task to the extreme. She has written a book called *The Power of Tests*, where she has looked into the consequences or impact of tests not only on societies, but also on individuals, and has probably become the most influential advocate of the critical perspective on the uses of language tests.

Shohamy claims that traditional testing is not interested in test use, in the motives for the introduction of tests, or in the consequences of tests and their effects on those who fail or succeed, considering tests as isolated events, detached from people, society, motives, intentions, uses, impacts, effects and consequences.

Shohamy's critical perspective: Tests emerging as power tools

According to Shohamy, tests emerged as a procedure for fair selection, based on the principles that they would: a) grant opportunities to all, regardless of people's background/origins; b) be objective; c) be scientific, applying methods for objectivity and fairness; and d) use objective item types, minimizing biases of judgment and reducing subjectivity with ratings. These principles would, according to the author, bring about fair selections, and would help turn ascribed systems/ascribed society – based on biased criteria like son-of-whom candidate, into achieved society – where all have the same equal chance of being selected.

The author claims, however, that the techniques developed based on those four principles – multiple choice items – turned out to be an illusion with respect to objectivity and fairness. The illusion of objectivity comes from the fact that black boxes could not be criticized. The illusion of fairness derives from the fact that the test writers used their knowledge as criteria for decision-making during the development of the test, imposing certain knowledge on the test takers.

Considering the consequences for individuals, and ultimately to society as a whole, Shohamy stresses that the power of tests is expressed when they, for example, force test takers into another profession, stigmatize people as failure, force people to choose between life and studying, cause individuals to have low trust in the themselves.

Tests are, in her view, seen as powerful, authoritative, frightening, deterring and controlling, leading necessarily to the detrimental consequences by which test takers see tests as powerful and themselves as powerless.

Still, according to the author, tests create winners and losers, successes and failures, rejections and acceptance. They can classify people, grant certificates and prizes, determine admission, decide on a profession, and turn friends or peers into enemies or rivals, especially in the case of fixed-quota openings, such as the university entrance examinations. In sum, tests can open or close doors.

Tests As Disciplinary Tools

Resorting mainly to Michael Foucault's (1979) ideas that society needs instruments of power and disciplinary tools, and that tests can be used as a controlling tool and as a disciplinary tool, Shohamy (2001) claims that as a controlling tool, tests have built-in features to be used for exercising power and control, such as hierarchy observance, people judgment, and establishment of the truth. As a disciplinary tool, they have the power to change behavior, dictate what to know, what to learn, what to teach, and cause fear and anxiety. Such power is also reflected in the fact that test demands are imposed from above and test takers are "forced into a position where they have no choice but to comply with these demands... without their voice being heard" (p. 19).

The author explains further six features of the power of tests. The first feature is that tests are administered by powerful institutions, which turn test takers into powerless individuals, since powerful institutions have the most power for decisions such as what and how to test, how to score, and interpret results, having total non-

negotiable control over most decisions which control and monitor the access to the desired object. The second feature is the use language of science, giving authority to the test and test results, since science is usually considered as objective, fair, true. The third feature is the use of the language of number, giving tests a symbol of objectivity, scientism, and rationalism.

The use of written communication is the fourth feature which she points out that affects negatively the degree of interaction, since it allows for little negotiation, correction, and argument. The fifth feature is the reliance on documentation by which the content of a test is recorded, making it possible to describe, analyze and compare individuals. Finally, the sixth feature, use of objective formats, allows for only one correct answer determined by the test writer in advance, thus, not open to interpretations.

In addition to these six built-in features, Shohamy points out that there are other features: tests a) are seen by the public as authoritative; b) allow flexible cutting score; c) are effective for control and redefinition of knowledge; d) have strong appeal to the public; e) are useful for delivering objective proof; f) allow cost-effective and efficient policy making; and g) provide those in authority with visibility and evidence of action.

She uses the descriptors suggested by Michael Foucault to further analyze the power tests have in our society. Thus, tests function as different acts on society such as: a) an act of surveillance by regulating behavior; b) an act of quantification by associating test takers with their scores; c) an act of classification by classifying students as success or failure; d) an act of standardizing populations by dictating real curriculum or *de facto* curriculum⁹; e) an act of judgment and sanctions through the use of records for the legitimacy of imposing sanctions; f) an act of demonstrating authority

⁹ The term *de facto* curriculum refers to what teachers actually use in classroom independently of the real curriculum, defined by authorities

by identifying authority, those making most decisions during the development, use or scoring of the tests; and g) an act of frightening and deterrence by having the instruments affecting everyone's lives in secret.

Shohamy proposes a scheme that describes the process underlying the power of tests. She believes that there are two components for the process of exercising power: a) the detrimental force of tests, i.e., high-stakes tests have a higher detrimental force; b) the features of the power of tests, such as the administration of the test by powerful organizations, the use of language of science and numbers, of written format, of objective formats, may legitimize the exercising of power.

These two components of tests have the power of changing behavior, exercised by those in power. Both test takers and their families will adapt to the demands of the tests, i.e., they will "comply with the demands of tests in order to maximize their score and gain the benefits associated with high scores" (p. 105). They will have more chances for success, and for the benefits involved such as study opportunities, job opportunities, better salaries, and recognition. Also, they avoid the bad consequences coming from the failure in the test.

Power and control can also be used as a way of implementing certain policies, that is, those in authority of making decisions, recognizing the power of tests, "take advantage of the phenomenon of the power of tests in order to change the behaviour of those affected by the tests, in line with certain agendas" (p. 106).

The Unchallenged Use Of Tests

According to Shohamy (2001), tests are used without much resistance from those affected, both directly and indirectly, since tests have enormous trust and support by the

public, and have become symbols of power since those submitted to it believe the power exists.

As to the support by the public, tests have become acceptable and recognized by everyone involved, since they all have some interest in tests: a) for the test takers, they provide recognition of their good qualities; b) for political institutions, they serve the cause of certain interest; c) for bureaucrats, administrators, and elite, they mean social order; d) for parents, they are an indication of what their children are able; e) for the elite, they are ways of control.

Concerning the mechanisms for symbolic power, still resorting to Michael Foucault (1979), Shohamy analyzes language tests as: a) contracts, through which groups cooperate with one another to maintain social order and the existing knowledge; b) rites of passage, since they are screening devices and maintain class differences; c) controlling and perpetuating knowledge, including knowledge socially recognized as legitimate; d) creation of dependence, since everyone is used to tests and to being evaluated through them; e) economic value in the form of certificates, promotions, etc; f) combination of language and tests – the power of knowing a language, in particular English, is combined with the power of tests; g) textual power, since written texts have power; and h) rituals with dates, times, people watching, score delivery, etc. Shohamy claims, however, that tests have turned from symbolic power to ideology, since they are believed by everyone, have their own rhetoric, myths, and numbers as a way of convincing people.

In sum, what seems to have been proposed as a solution for assessment procedures based on merit has also had its share of criticisms. In particular, two aspects may be mentioned. One is that merit is a construct which is defined by people, according to their values and interests. The other is that tests are not an entity in itself,

but reflect the contradictions of the society which uses them, a society which needs instruments for control and discipline. My analysis of the university entrance examinations includes this perspective, since they are the tests with the highest stakes among the tests analyzed, hence with the highest impact on both Brazilian society and individuals.

CHAPTER III

Defining the Construct for the Investigation of Validity

In this chapter, I aim to present the review of literature on the constructs used in this research: language ability, and reading ability in particular. Ability, according to Howe (1996), may be defined in two different ways: a) purely descriptive, referring to what a person can do, i.e., a person has an ability means that the person can do something; and b) explanatory, referring to what explains why a person can do something, i.e., what a person has that allows him/her to accomplish something, the underlying reasons for the person's success.

Following these two conceptions of ability, it is possible to understand the distinction suggested by Bachman (1990) to define the construct of language ability as follows: a) a real-life approach, concerned with identifying a domain of actual use to be characteristic of performance of competent language users, for example, the American Council on Teaching of Foreign Languages (ACTFL) rating scales; and b) an interactional/ability approach, defined in terms of component abilities, such as the communicative frameworks proposed since the 1970s.

Although the real-life approach of the ACTFL proficiency rating scale has had great influence on testing practice (Brindley, 1998), it will not be used in this research as part of the analysis due to problems and uncertainties around it (Bernhardt, 1986, 1991; Brindley, 1998; Chastain, 1989; Byrnes, 1986).

I, thus, present and discuss, in section 3.1, Bachman (1990) and Bachman and Palmer's (1996) framework of language ability as an example of the explanatory approach for language ability. In section 3.2, I present accounts, with an explanatory

approach, of the reading ability in terms of its underlying processes, of its skills and subskills, of the types of reading determining the specific skills for global and local comprehension, and of the resulting mental representations. In section 3.3, I present the definition of the construct possible to be assessed in the testing situations. In section 3.4, I present a discussion on the correspondence of item used and skill assessed.

3.1 – Framework by Bachman: An Explanatory Approach

Bachman's (1990) language ability framework is theoretical, based on the ongoing debate of what communicative language ability (CLA) is. In general, the framework proposes that communicative language ability consists of both knowledge, or competence, and the capacity for implementing or executing that competence in appropriate, contextualized communicative language use. It claims that knowledge structures (of the world) and language competence (knowledge of language) are fed into strategic competence.

Bachman (1990) discusses the three most essential components concerning specifically language, namely, language competence or language knowledge, strategic competence as metacognitive strategies, and psychophysiological mechanisms, focusing only on the two first components, since "it is this combination of language knowledge and metacognitive strategies that provides language users with the ability, or capacity, to create and interpret discourse, either in responding to tasks on language tests or in non-test language use" (Bachman & Palmer, 1996, p. 67).

I will, then, first, present Bachman's framework of language ability, proposed by Bachman (1990) and repeated in mostly the same way in Bachman and Palmer (1996). Next, I will be presenting Bachman's (1990) definition of strategic competence and

explanation of its role in language performance, as well as the somewhat different definitions of strategic competence and explanation of its role in language performance put forth by Bachman and Palmer (1996).

Bachman's framework

Bachman (1990) presents a framework with an explanation of language competence or knowledge, which is a reinterpretation and/or an expansion of other models of communicative competence previously proposed, such as Canale and Swain's (1980) model. Bachman's description of language competence within his framework is also based on the empirical findings by Bachman and Palmer (1982), whose results showed the competences closely associated with each.

According to the author, language ability is comprised of organizational knowledge and pragmatic knowledge. Table 4 below, presented in Bachman and Palmer (1996), summarizes the components of language ability as considered by Bachman (1990).

Table 4: Areas of language knowledge, as presented in Bachman and Palmer (1996).

<p>Organizational Knowledge¹⁰ (how utterances or sentences and texts are organized)</p> <p>Grammatical Knowledge (how individual utterances or sentences are organized)</p> <p>Knowledge of vocabulary Knowledge of morphology Knowledge of syntax Knowledge of phonology/graphology</p> <p>Textual Knowledge (how utterances or sentences are organized to form texts)</p> <p>Knowledge of cohesion Knowledge of rhetorical organization</p> <p>Pragmatic Knowledge (how utterances or sentences and texts are related to the communicative goals of the language user and to the features of the language use setting)</p> <p>Functional¹¹ Knowledge (how utterances or sentences and texts are related to the communicative goals of users)</p> <p>Knowledge of ideational functions Knowledge of manipulative functions Knowledge of heuristic functions Knowledge of imaginative functions</p> <p>Sociolinguistic Knowledge (how utterances or sentences and texts are related to the features of language use setting)</p> <p>Knowledge¹² of dialect or variety Knowledge of differences in register Knowledge of naturalness Knowledge of cultural references and figures of speech</p>
--

¹⁰ Bachman (1990) uses the word *competence* rather than the word *knowledge* for all the types of knowledge involved in language ability.

¹¹ This knowledge is called illocutionary rather than functional in Bachman (1990, p. 87).

¹² Bachman (1990) uses sensitivity to rather than knowledge of for all the types of knowledge under sociolinguistic knowledge.

Assuming that all the various types of component knowledge integrating Bachman's (1990) model are widely known in the field of second language acquisition (SLA) studies, they will be briefly presented here.

Organizational knowledge is involved in the controlling of the production or recognition of grammatically correct sentences, as well as the understanding of their propositional contents. It comprises grammatical knowledge and textual knowledge.

Grammatical knowledge consists of knowledge of vocabulary, morphology, syntax, and phonology/graphology, that is, the areas of language knowledge drawn upon to produce or understand formally accurate utterances or sentences. Textual knowledge consists of two distinct areas of knowledge: knowledge of cohesion and knowledge of rhetorical or conversational organization. The former is involved in the production or comprehension of explicitly marked relationships among sentences or utterances, whereas the latter is involved in the comprehension of organizational development in written texts or in conversations.

Pragmatic knowledge refers to the ability to relate utterances or sentences and texts to their meanings in the context, and to the intentions of the users, and it involves two areas of knowledge: functional knowledge and sociolinguistic knowledge. Functional knowledge makes it possible for the language users to interpret the intentions underlying the use of language within their context, and includes four categories of language functions:

- Knowledge of ideational functions, including language use to express or comprehend ideas, knowledge, or feelings;
- Knowledge of manipulative functions, including: a) instrumental functions - to have people do something, e.g. request, suggest, warn; b) regulatory functions -

used to control people's behavior; c) interpersonal¹³ functions - used in interpersonal relationships, as greetings, compliments, etc.

- Knowledge of heuristic functions, allowing language use to extend language use for teaching and learning, problem-solving, and retention of information.
- Knowledge of imaginative functions, involving the use of language for humor or esthetic purposes, as in jokes or figurative language or poetry.

Sociolinguistic knowledge involves the knowledge of the conventions of the language which allow the user to make the most appropriate use of figures of speech, expressions, dialects or varieties, and cultural reference.

Strategic competence is central in his framework. Strategic competence is responsible for implementing "the components of language competence in contextualized communicative language use" (Bachman, 1990, p. 84), thus responsible for relating language competencies to context of situation and to knowledge structure.

Bachman wants to stress that since communication is an interchange between context and discourse, and interpretation requires the use of available language competencies for relevant information and its matching with information in discourse, it is the role of strategic competence to "match the new information to be processed with relevant information available and map this onto the maximally efficient use of existing language abilities" (1990, p. 102). The author, then, does not accept the notion that strategic competence is only called upon when there is a problem in communication to be compensated by other means, and considers the role of strategic competence in the interaction between the various competencies and the language use context.

¹³ Interpersonal is called interactional function in Bachman (1990).

Bachman's description of strategic competence in communicative language use, he claims, is an extension of Faerch and Kasper's (1983) formulation, and the three components are: assessment, planning, and execution. The assessment component is responsible for the communicative goal in relation to the context and the interlocutor. More specifically, it allows the language user to a) identify the information needed for the communicative goal in the context; b) determine the resources available to achieve the goal; c) use language and knowledge shared with the interlocutor; and d) evaluate whether or not the communicative goal was accomplished.

The planning component is more related to the linguistic resources, and its role is to find relevant items from language competence, for example the appropriate forms of address and questioning routines, and make a plan for a communicative goal. The execution component involves considering the plan in the modality and channel appropriate to the context and communicative goal, and drawing on the relevant psychophysiological mechanisms¹⁴ to implement the plan.

There are, however, other factors influencing performance which should be taken into consideration. In addition to the components of language competence already integrated in the model of language ability proposed by Bachman (1990), Bachman and Palmer (1996) claim that, for language use, there are other components to be considered: topical knowledge, personal characteristics, and affective factors (called by them affective schemata).

Since these factors affect performance, the authors claim that they must be carefully considered in the design, development and use of language tests. They explain the factors as follows. Personal characteristics refer to factors such as age, sex,

¹⁴ Bachman (1990) defines the psychophysiological mechanisms as: in the receptive language use "auditory and visual skills are employed, while in productive use the neuromuscular skills (for example, articulatory and digital) are employed" (p. 107).

nationality, resident status, native language, level and type of general education, and types and amount of preparation or prior experience with a given test which may influence performance on language test.

Topical knowledge refers to the knowledge structures stored in the long-term memory, used to enable language users to refer to their world. Affective schemata are related to the characteristics of language use tasks and the past emotional experiences in similar contexts, and determine greatly the affective responses to a particular task. It is important to note that, in including affect in their account of language use, they make it clear that it is a crucial component in facilitating or limiting the flexibility of the language users' use of the language, that is, their performance.

According to their model, affective schemata affects, one way or the other, strategic competence and, consequently, the three general areas in which its metacognitive components operate for both language use tasks and language test tasks: goal-setting, assessment, and planning.

Goal setting can be roughly defined as deciding what one is going to do, which requires identifying the language tasks, choosing the tasks, and deciding whether to try them or not. Flexibility is the aspect affected by affective schemata, and in the case of a testing situation, it is very limited by the tasks provided in the tests. Assessment allows the language user to relate his/her topical knowledge and language knowledge to the language use situation or testing situation, considering the affective responses.

As the third metacognitive component, planning is related to the use of language knowledge, topical knowledge and affective schemata, and how they will be best combined in response to tasks, involving the aspects of selecting from the topical and language knowledge for the context, formulating the plan to be carried out in response to the task, and selecting a plan to be carried out in response to the task.

In addition to providing an account for the understanding of language ability, what is specifically relevant in their framework for the present research is that language ability is seen as strategic, and the language user seen as making use of the available competencies for the communicative goal, as discussed previously. Also relevant is that the authors incorporate the idea of affective schemata as influencing the use of language, in particular the strategic competence involved in language use. In terms of their notion of authenticity, as seen in chapter 2, and the need for a high degree of correspondence between the test tasks and target language use (TLU) tasks, it is possible to argue that the correspondence of non-cognitive aspects such as the characteristics of the setting (physical characteristics, participants, and time of the task), which load affective schemata, must be high too, otherwise the use of language will be affected negatively. This is further discussed in section 4.7.

3.2 – Reading Ability: Accounts within the Explanatory Approach

Reading has been explained as comprising of two distinct, though interactive, knowledge bases which have been described as declarative and procedural¹⁵ (Just & Carpenter, 1984). Gagné, Yekovich, and Yekovich (1993) define reading as declarative knowledge as referring to conceptual knowledge, i.e., knowledge of the topics mentioned in the text, text schemas, and vocabulary, and define reading as procedural knowledge as including the underlying skills forming the basis of the conceptual knowledge, involving the following component processes: decoding, literal comprehension, inferential comprehension, and comprehension monitoring. Due to

¹⁵ Declarative knowledge has been defined as the knowledge base of ‘knowing that...’ and procedural knowledge as “the knowledge used in performing actions, including mental actions (Just & Carpenter, 1984).

their relevance for the definition of the construct of reading ability, decoding, literal comprehension and inferential comprehension are explained below.

Decoding refers to the cracking of the code to make it meaningful, i.e., the process through which the reader recognizes the sight vocabulary and activates its meaning in the long-term memory. The decoding process stimulates the literal comprehension process, whose function is to derive literal meaning from the print, and which encompasses two processes: a) lexical access, responsible for identifying and selecting the appropriate meaning of the words; and b) parsing, which uses the syntactic and linguistic rules of the language to derive meaning from larger units of meaning, such as the meaning of a phrase, a clause or a sentence. These processes are considered lower-level processes since they are based on basic linguistic knowledge, and on the identification of the information literally given in writing, thus explicitly present in the text.

Deeper and broader understanding of ideas is possible through inferential comprehension, which includes the following processes: integration, summarization and elaboration. The integration process is the result of a coherent representation of the ideas in the text, i.e., how the reader connects the propositions together. Summarization is the mental outline of the hierarchically arranged propositions that capture the main ideas of the text. Elaboration on the propositions is when the readers bring their prior knowledge to the new information presented in the text in their pursuit of acquiring declarative knowledge in terms of the knowledge of the subject matter. These processes imply drawing inferences, based on background knowledge, to what is only implicitly given in the text, and are, thus, considered higher-level processes.

The three inferential processes - integration, summarization, and elaboration, however, involve background knowledge in two different levels. Both integration and

summarization processes require inferences necessary for comprehension since their role is to "organize new information by building a coherent meaning representation" (Gagné et al, 1993, p. 278). Elaboration, on the other hand, relies on inferences brought to bear which are not essential for the organization and coherence building, thereby not being essential for comprehension to occur, but rather, an addition of "new ideas gleaned from the text" (p. 278).

Reading may also be defined in terms of skills and tasks involved. Nuttall (1996) explains reading as involving lower level processing called *reading for plain sense* and higher level processing called *understanding discourse*. Reading for plain sense involves, according to the author, three reading skills requiring mostly bottom-up processing: understanding syntax, recognizing and interpreting cohesive devices, and interpreting discourse markers.

The tasks involved in understanding syntax are: a) identification of cohesive elements (further explained below); b) identification and understanding of coordinating conjunctions; c) recognition and understanding of the constituents of a noun phrase; d) recognition and understanding of nominalization; e) identification of the main verb and other finite verbs, as well as the subject and object, and the establishment of the boundaries of each clause; and f) understanding of participles and infinitives in non-finite clauses as in the example *smoking is bad for your health*.

Nuttall (1996) defines cohesive devices as those devices the reader must use when making the expected connections between the ideas expressed in a text, when identifying the cohesion of the text. They are mainly concerned with the signification of the text, and identification of references for the reader to be able to understand the plain sense of the text (p. 94), and can be divided into three subclasses, reflecting three

different subskills necessary for interpreting discourse: a) interpreting pro-forms; b) interpreting elliptical expressions; and c) establishing lexical cohesion.

Pro-forms are words used in texts to avoid repetition: words such as *it, our, this, one, so/not* (as in I think so/not), and comparatives (smaller, same, other). Elliptical expression, or ellipsis, is the piece of information in the text which is not present in order to avoid unnecessary repetition, and must be provided by reference to other information previously stated in the text.

Lexical cohesion must be established by the reader, that is, the reader must be able to interpret the relationship between a lexical item and other parts of the discourse directed by the writer. Examples of such case are synonymy, hyponymy, metaphor, text-structuring words, and pin-down words.

While it is the case that synonymy, hyponymy, and metaphor are familiar to most people in the area of linguistics and applied linguistics, thus not demanding explanation, text-structuring words and pin-down words deserve some explanation.

Text-structuring words are words to be lexicalized within its context, that is, the reader must infer its meaning by reference to some information usually previously stated. Examples are: *issue, methods, events, views, explanations, and phenomena*. Pin-down words refer to propositions, thus carrying their underlying propositional meaning. The example given by the authors is the word *approach*.

Discourse markers signal relationships between different parts of the discourse, and are used by the text writer to show his/her intended relationships between the parts of the text. They may indicate the functional value of a sentence. They are markers that signal sequence of events, discourse organization, and the writer's point of view.

Reading into discourse is concerned with what the writers mean by what they say, that is, what is either presupposed or implied by the writer. The author claims that five

skills are essential for reading into discourse. The first skill, the recognition of the functional value, may be the result of identification of the signaled discourse markers, or the result of the inference process in an attempt to understand the intended message when it is only implicit. The second skill, recognizing the text organization, the reader must be able to recognize the rhetorical organization of the text to be able to interpret the text, in this case, how the ideas are related to one another, which will facilitate comprehension.

The third skill, recognizing presuppositions, is based on the principle of efficient communication, that is, the writer will not include in the text what may be presupposed for the reader. The writer will select some information to include in the development of the text as well as some information to be left out, usually part of the shared knowledge and experience, and part of the shared opinions, attitudes and emotions, even though sometimes constituent of the line of reasoning or argument.

The fourth skill, recognizing implications and making inferences, is also based on the principle of efficient communication. The writer will not include in the text, but will imply, the information he/she expects the reader to be able to infer, and, the reader must be able to reconstruct the writer's unstated presuppositions and draw certain unstated conclusions from facts, or points in argument, using the evidence provided by the writer.

The fifth skill, prediction, refers to the prediction of the text content based on the title, leading to expectations of the text content. Prediction, based on the world knowledge, will assist the comprehension of the incoming information, since this information can be framed, i.e. can be "fitted into the existing framework of ideas" (p. 118).

Reading can also be defined in terms of the types of reading determining the specific skills for global and local comprehension¹⁶. Urquhart and Weir (1998) claim that there are specific skills for both global and local comprehension for two types of reading. For expeditious reading, specific type for leisure reading, skills for global comprehension are skimming for topic and main ideas, and search reading to locate quickly and understand relevant information. The skill for local comprehension is scanning to locate specific information.

For careful reading, specific type for studying, the skills for global comprehension is reading carefully to establish accurate comprehension of the explicitly stated main ideas the author wishes to convey, i.e., propositional inferencing. Skills for local comprehension are understanding syntactic structure of sentence and clause, understanding lexical and/or grammatical cohesion, and understanding lexis/deducing meaning of lexical items from morphology and context.

Reading may also be defined in terms of the resulting mental representation of the text, which most theorists in the area of language processing agree today to be the result of some combination of text-derived and knowledge-derived information, or text-based and reader-driven information. Theories of discourse processing today, mostly based on the works by Kintsch and van Dijk (1978), van Dijk and Kintsch (1983), have postulated that there are two levels of representation resulting from the reading of the text. They are called textbase model and situation model.

The textbase model refers to “those elements and relations that are directly derived from the text itself” (Kintsch, 1998, p. 103). The mental representation

¹⁶ Global comprehension refers to “the understanding of propositions beyond level of microstructure, that is, any macroproposition in the macrostructure, including main ideas and important details [and] local comprehension refers to the understanding of propositions at the level of microstructure, i.e., the meaning of lexical items, pronominal reference, etc.” (Weir et al, 2000, p. 23-4).

resulting from this level of comprehension refers mostly to the propositional network of the text, representing its meaning (Kintsch, 1998, p. 105). The situation model is a construction of a coherent structure of the text with the establishment of the necessary links and integration with prior knowledge, all based on the knowledge of the language, of the world, of the communicative situation, and on personal experience. It is “a construction that integrates the text-base and relevant aspects of the comprehender’s knowledge” (Kintsch, 1998, p. 107).

In sum, reading may be accounted for in many ways. In reviewing the research on the accounts of reading, Grabe (1999) concludes that, although there are some disagreement as to some specific mechanisms involved in reading comprehension, there is the agreement that there are two networks of comprehension, one, called text model, reflecting the information presented in the text, and the other, called situation model, including “much more reader background knowledge, affective responses, and individual interpretations of the text information” (p. 17).

According to Grabe (1999), text model is a close representation of comprehension up to the level extending beyond the sentence-level propositional integration, the textual propositional network, which is based only on information presented in the text, or information needed for the integration of the propositions. This model is hierarchical and gradually includes higher-level macropropositions with the main ideas of the text, based on the information added through lower-level processing, but selected for further restructuring based on the reader’s prior knowledge and goals.

The situation model representing an interpretation of the text information is constructed along with the text model, involves reader’s background knowledge to a further extent, and is based on the reader’s goal for reading, motivation, attitudes and evaluations of the information in the text.

Grabe claims that this account of comprehension as text model and situation model allows to see the reader as both understanding the text and interpreting the text, i.e., being able to “both recognize and understand the information in the text, and also to create an interpretation that is unique to the particular reader” (p. 19). The reader, then, can be seen as providing similar summaries, and also distinct, based on their own background knowledge and interest.

The recognition of the contribution of individual factors such as background knowledge for comprehension or interpretation imposes a problem for testing: who has the right answer as to the mental representation of the text, the test user or the test taker? Or rather, can there be right answers, especially in the case of open-ended questions?

Also, the recognition of the contribution of the individual factors imposes a problem for construct validity: what is being assessed in a reading test, reading ability or background knowledge? What is the variable being assessed? If an item can be gotten right with the required knowledge, but cannot without it, it is plausible to argue that the variance being assessed is background knowledge.

Furthermore, the recognition of the contribution of the individual factors imposes a problem for consequential validity: is it fair to use items whose answers presuppose knowledge which may not be shared by some test takers or some groups of test takers? Is it fair to go against one of the principles¹⁷ of good item writing which says that items must be developed to allow for the best performance of test takers so that language ability assessed can be demonstrated? As Bachman and Palmer (1996) stress, “certain test tasks that presuppose cultural or topical knowledge on the part of the test takers

¹⁷ Principle also shared by Bachman and Palmer (1996) in their philosophies of good item writing.

may be easier for those who have that knowledge and more difficult for those who do not” (p. 65).

In sum, there seem to be methodological and ethical (fairness) problems concerning background knowledge and testing and, for the purpose of testing, it may be a factor to be controlled. As Alderson (2000) explains, “it [background knowledge] is therefore a candidate for the sort of variable we would wish to control or neutralise and every attempt should be made to allow background knowledge to facilitate performance rather than its absence to inhibit” (p. 121).

3.3 – Defining the Scope of the Construct to be used for the Analysis

Defining the scope of a construct implies making decisions as to what factors must be included and also what factors must not be included. This definition helps in the identification of the factors measured, with the consequent interpretation that other factors are source of invalidity. As Bachman (1990) points out, the interpretation of the effects of factors as source of measurement error, test bias, or as part of the language abilities intended “will depend on how we define the abilities and what use is to be made of test scores in any given testing situation” (p. 278).

In addition to the methodological and fairness problems in relation to the factor of the background knowledge, as pointed out in the previous section, there is also the factor of the adequacy of the construct to the characteristics of the examinations and of test takers’ expected background knowledge.

Bachman and Palmer (1996) offer three options for defining the scope of the construct in relation to topical knowledge¹⁸, considering the characteristics of the examinations and of the test takers' expected knowledge. The options are: 1) construct definition including only language ability, not including topical knowledge; 2) construct definition including topical knowledge; and 3) separate constructs defined for language ability and topical knowledge

In the first case, construct definition including only language ability, the inference to be made is related only to components of language ability. The typical situation is language programs, and academic and professional programs, where language ability is important for the decision based on the test, and test takers are heterogeneous in relation to topical knowledge.

In the second case, construct definition including topical knowledge, the inference to be made is related to the ability to process specific topical information through language. The typical situation refers to groups with homogenous topical knowledge, such as specific purpose courses, where language and topical information are part of the course syllabus.

In the third case, language ability and topical knowledge defined as separate constructs, the inference to be made refers to components of language ability and areas of topical knowledge. The typical situation is when test developers want to measure both language ability and topical knowledge, but do not know whether the test takers have homogenous topical knowledge or not.

Concerning the specific factor of topical knowledge, the three testing situations analyzed in this research – proficiency examinations, university entrance examinations, and EAP tests for heterogeneous groups – should fall into the first case of assessing

¹⁸ The definition of topical knowledge by the authors is that it refers to knowledge schemata or real-world knowledge (p. 65).

language ability as the construct, due to their characteristics of dealing with heterogeneous groups in terms of background knowledge.

There are, thus, methodological and fairness reasons, as well as reasons related to the characteristics of the examinations and of the test takers' expected background knowledge, to control for the factor of background knowledge and not include it in the construct of reading in this research. I will, then, analyze the test items following this definition of the construct. Anything other than language ability will, thus, be considered source of invalidity evidence.

Once the construct has been defined as language ability, more specifically reading ability, items must be developed to reflect the choice in a way to minimize, for the purpose of testing, the effects of background knowledge for the responses to the items. Two ways of minimizing its effects, suggested in the literature, are, according to Urquhart and Weir (1998), by choice of text or choice of task.

As for the first choice, three possibilities have been suggested: 1) using a variety of short texts with a wide range of topics; 2) using texts unfamiliar to all candidates so that "text variables rather than background knowledge have the most influence" (Urquhart & Weir, 1998, p. 116); and 3) using texts whose topic familiarity is established in advance through interviews or questionnaires with the candidates. As for the second choice, task choice, the authors, recognizing the difficulty of knowing the level of inferences involved, suggest that test items should focus on propositional inferences rather than on pragmatic inferences.

Hughes (2003) contributes to the distinction between the two types of inferences by providing the following definitions: propositional inferences are "those which do not depend on information from outside the text [...] [and] pragmatic inferences are those

where we have to combine information from the text with knowledge from outside the text” (p. 139).

The example given by the author for propositional inference is that it is possible to infer that Harry was working at her studies based on the information *Harry worked as hard as she had ever done in her life. When the exam results came out, nobody was surprised that she came top of the class.* The example given for pragmatic inference is that it is impossible to infer if some drivers drove fast or slowly with the information *it took them twenty minutes by road to get from Reading to Heathrow airport,* unless the reader knows the distance between Reading and Heathrow. It can be inferred that they drove fast if it is known that the places are very distant. It can be inferred that they drove slowly if it is known that the places are close to each other. If this specific information about the distance is not in the text, and is not part of the reader’s background knowledge, such inferences are not possible.

As to the choice of task, a distinction has also been made by Pearson and Johnson (1978). The authors have characterized the relationship of a passage and the questions developed for it, that is, the types of questions and level of comprehension required by them. They are: textually-explicit questions, textually-implicit questions, and scriptally-implicit questions.

Textually-explicit questions are those whose relation is based on the text, that is, both questions and answers are based on the text, but at the same time, restricted to the information explicitly given in the text. This kind of question-answer relation happens “when both question and answer are derivable from the text and if the relation between question and answer was explicitly cued by the language of the text” (p. 163). According to the authors, literal comprehension is involved for the answer, which is also called ‘reading the lines’.

Textually-implicit questions refer to the relation question-answer based on the text and only implicitly given in the text, i.e., “if both question and answer are derivable from the text *but* there is no logical or grammatical cue tying the question to the answer *and* the answer given is plausible in light of the question” (p. 163). According to the authors, inferential comprehension is involved here, which is also called ‘reading between the lines’.

For this kind of questions, the plausibility criterion is what distinguishes plausible inferences from textual intrusions, the latter characterized as coming from the text, but with no argument for them to be considered plausible answers to the questions. Divergent responses differ from textual intrusions since they are based on some logic that can be recoverable from plausible inferences drawn by the reader, i.e., “the logic is plausible” (p. 164).

Scriptally-implicit questions occur when the question is derivable from the text, but a plausible nontextual response is to be given (p. 164). Since the answer cannot be found in the text itself, it requires higher contribution from the reader’s background knowledge, ‘reader’s fund of knowledge’ as called Pearson and Johnson (1978). It can be argued that this type of questions may not be comprehension questions at all, since “they rely on information outside the text” (Alderson, 2000, p. 87). In fact, based on my analysis of the question-answer relation of the examples given by the authors, they cannot, in my view, be called comprehension questions.

Although this three-level classification has been widely used, either with Pearson and Johnson’s (1978) ‘textually-explicit questions’, ‘textually-implicit questions’, and ‘scriptally-implicit questions’, or with the more general ‘reading the line’, ‘reading between the lines’, and ‘reading beyond the lines’, a somewhat different level may be added. In presenting her six-level classification, Nuttall (1996) contributes with the

type of question which involves further processing than literal comprehension, since it requires elementary inferences for the reorganization or reinterpretation of the information present in the text, type that she refers to as ‘questions of reorganization or reinterpretation’.

Pearson and Johnson’s (1978) typology, as well as Nuttall’s (1996) type of question focusing elementary inferencing will be used for analysis to characterize the source of information required by each item.

3.4 – Discussing the Correspondence of Item used Skill assessed

The idea of choosing tasks to assess certain or intended skills or levels addresses a very controversial issue of whether it is possible to say that one question or item assesses specifically the skills or levels intended. This raises the broader issue of whether or not reading can be assessed as smaller skills or components skills or in its various levels, which relates to the ongoing debate on the two alternative views of reading: reading as a unitary ability, or as a multidivisible ability. The view of reading has a direct influence on the type of testing chosen: the view of reading as a unitary ability will favor integrative testing, and the items will be chosen to assess global comprehension, whereas the view of reading as a multidivisible ability will allow for discrete-point testing.

Following the view that reading is a unitary ability, Bernhardt, (1991) defines reading comprehension as a “constructive construct – not one that is a sum of a number of discrete points” (p. 193), and claims that assessment instruments must be

integrative¹⁹ in nature and examine if a text actually communicates a coherent message to the reader. Also inclined to accept this view, Alderson (1996) argues, “it may simply be enough to determine whether or not a student has understood a text” (p. 220).

This unitary competence hypothesis (Hughes, 1989) or unitary trait hypothesis (Brown, 1987), proposed by John Oller in the 1970s, suggests the indivisible view of language, based on which language ability cannot be divided into separable components to be tested, that is, that the "nature of language ability was such that it was impossible to break it down into component parts" (Hughes, 1989, p. 62), meaning that discrete-point testing does not show much about the overall ability. However, language competence was proven less likely to be assessed as a unitary ability, and the unitary hypothesis was abandoned as an account of language competence (Brown, 1987), creating a vacuum in terms of theoretical issues for language testing (Bachman, 1989). Brindley (2001) adds that Oller himself, in the end, accepted the idea of multiple components (p. 139).

Although recognizing the integrative nature of reading, i.e., that the sum of the parts may not necessarily equate comprehension, Urquhart and Weir (1998) claim that it is “difficult to maintain that reading is a unitary ability” (p. 128), thus favoring the concept of reading as multidivisible²⁰, that is, reading is made up of various components, such as word recognition. Moreover, the authors stress that test developers do focus on the individual reading components when constructing test items, considering that, for the purpose of teaching and testing, reading can be broken down into underlying skills or abilities.

¹⁹ Integrative test is defined in opposition to discrete-point test, where the former refers to combining many language elements to complete a task, and the latter to the testing of each element at a time (Hughes, 1989).

²⁰ Although favoring the concept of multidivisible ability, the authors stress that it is dangerous to date to claim that there is such a purely multidivisible, since there are people who can establish the macrostructure of the text without knowing the meaning of some words (p. 140).

The debate on the unitary versus multidivisible hypothesis has two related questions: 1) if there are skills which can be described separately and what they are; and 2) if there is a hierarchy or implicational relation among them, that is, if higher order skills subsume lower order skills.

Alderson and Lukmani (1989) published the results of their research investigating whether reading skills can be identified separately. Based on previous studies which concluded that the participants were able to answer higher-order questions but failed to answer correctly lower-order questions, they hypothesized that there was no hierarchy involving lower-order skills and higher-order skills, where the latter would depend on the former. The general conclusion, according to the researchers, was that “there is no implicational scale in reading in a second language such that one needs lower order abilities before one can progress to higher order questions” (p. 269).

Recognizing that the results of their research produced “only very tentative conclusions” (Alderson, 1990a, p. 428), Alderson decided for a continuation of the study. Alderson (1990a) carried out a research to investigate the existence of separate skills and hierarchy among skills. The results showed that there was no agreement among the ‘judges’ for some items, and very little agreement for the intended skills, leading to the conclusion that the skills used are interrelated and not discrete, that it is not possible to claim there is a hierarchy among the skills, and that it is not possible to say that one item is testing one skill. The researcher, thus, challenges the separation of skills and the hierarchy of skills, where the higher order skills subsume the lower order ones.

The publication of these results caused the debate around the reading construct and the correspondence of item and skill assessed to be polarized. Weir, Hughes and Porter (1990) published an article challenging the results of Alderson’s (1990a)

research. Recognizing that this kind of conclusion would have serious implications for the development of valid tests in reading, the authors state that the article requires careful reading, since there was “mistaken thinking at a number of crucial points in the article, weakness in the methodology, and a number of inaccuracies” (p. 505).

The authors stress that, with respect to the methodology, the expert teachers/judges were not reliable judges for the judgment and received little training²¹; with respect to the inaccuracies, the ‘judges’ did not have clear definition as to what higher- and lower- order skills were; and with respect to the mistaken thinking, higher- and lower- order skills cannot be judged in relation to the difficulty indices calculated based on the difficulty to answer test items, as was concluded by Alderson in his research.

In a follow-up research, part of his same project, Alderson (1990b) investigated the validity of the correspondence of method/item used and the skill assessed by analyzing the reading skills used to solve test items. The results led the researcher to the conclusion that “... test-taking process (and therefore, by inference, at least part of the reading process) probably involves the simultaneous and variable use of different, and overlapping ‘skills’” (p. 478), which, in his view, would be evidence enough to challenge the commonly accepted idea that different aspects of reading can be assessed through different methods or items.

As in the previous research, I believe that this one has its own problems. First, the research involved only two participants, with only 10 items used for data collection. Second, the data were collected through introspection and retrospection, using the target language, of which the participants were only learners. As all the conclusion of

²¹ Alderson (2000) recognizes the fact that the judges were not trained and argues that it was part of the methodology not to train the ‘judges’, since “such training would amount to cloning” (p. 96), that is, every judge trained to do and see the same thing.

the research is based on these reports, there seems to be, also here, methodological fragility.

In addition to this methodological fragility, Alderson extends to reading his conclusions about processes to solve test items. There are a number of factors making up the process of solving test items, called testwise strategies, which are not necessarily part of the reading process. The challenging of established presuppositions, as claimed the researcher, requires, in my view, research with less methodological problems.

In a very extensive research on the subject, Bachman, Davidson, Ryan, and Choi (1995) conclude that it is possible to get high agreement on what an item is testing when expert judges are involved. The authors attribute this consistency to the rating instruments used, which made it possible for the raters to “focus almost microscopically on very specific aspects of content... and provided a fixed range of judgments, as indicated in the rating scales for the various facets” (p. 122).

More recently, Alderson maintains a similar position with respect to the issue of the correspondence of method/item and skills assessed. Alderson, Clapham and Wall (1995) note that the understanding of such relationship is still “so rudimentary that it is impossible to recommend particular methods for testing particular language abilities” (p. 45), and comment that finding this correspondence would be compared to finding the ‘Holy Grail’ of language testing.

Despite the problems with his research as aforementioned, Alderson may have raised a critical issue with respect to the correspondence method/item and skill assessed when he claims, “answering a test question is likely to involve a variety of interrelated skills, rather than one skill only, or even mainly” (Alderson, 1990a, p. 436).

Skills such as identifying main idea, understanding explicit or implicit information, or understanding the relations within sentences or across sentences, may

be the ones assessed for the inference of the reading competence of some test taker. In case the test takers answer the item correctly, they will have this skill as part of the characterization of their competence. However, as research on testwise techniques shows, test takers may get an item correct without having the skill. Conversely, test takers may have the skill without being able to show it, due to poor linguistic knowledge, and, as a consequence, have a mistaken characterization of their reading competence.

Some studies show that reading skills may not be used in reading in a second language due to poor linguistic knowledge, requiring a linguistic threshold level to be used. Alderson (1984) concluded that readers in a foreign language cannot use semantic constraints provided by the context until they have reached a threshold level in the language. Carrell (1991) concluded that a threshold level is necessary for the readers to make inferences and identify the authors' position. Zwann and Brown (1996) concluded that integrative processes, necessary for the integration of information across sentences and, hence, necessary for the construction of coherence in a text, are affected negatively when linguistic knowledge does not allow for efficient syntactic and lexical processing.

Therefore, it is possible to argue that some test items may not be assessing the intended skills, and may be assessing, ultimately, the linguistic knowledge of the test takers, particularly of those who have not reached the linguistic threshold level to read the target texts.

Seeking to investigate the important issues in reading assessment discussed by Alderson and Lukmani (1989), Alderson (1990 a, 1990b), and mostly accepted by Alderson et al (1995) and Alderson (2000), Weir, Huizhong and Yan (2000) carried out a very extensive and detailed research. Their main theoretical objectives were to solve

the issue of whether reading is a unitary activity or whether it is made up of separable components, as well as to provide evidence of the correspondence of test method/item used skill/strategy assessed. That is, an investigation into the universe of the unidimensionality versus divisibility of the reading construct and the relative contribution of different parts to the reading ability.

The quantitative study involving different statistical analyses showed that the components in the test were measuring different parts of the reading construct, thus pointing to at least a bi-divisible view of reading, with vocabulary loading as a separate factor, in addition to general reading comprehension.

The qualitative study involved collecting data through: a) EAP reading experts' judgment on the skills and strategies tested - language testing experts and reading experts with their professional opinion of the constructs; b) students' introspection on the process of taking the test; and c) students' perceptions (retrospection) of the test conditions and the skills and strategies tested - a checklist for students to tick after finishing each section of the test.

The results of the qualitative study based on the experts' judgment showed the percentage of respondents who agreed with the developers' view of the primary focus of each section ranges from the lowest 88% (for skimming and search reading sections) to the highest 100% (for the careful reading sections), thus confirming the test developers' expectations. The researchers concluded that experts' judgments confirmed their expectations in terms of the skill/strategy being tested as a primary skill in each section. They also showed that the respondents identified secondary skills, suggesting that other skills/strategies than the intended ones had been used.

The retrospection study aimed at investigating the students' perceptions of the skills and strategies used while taking the test, confirmed their expectations to a lesser

degree than the experts' judgment. Since one possible explanation for the lower agreement of students' perception with their expectations and with the experts' judgment was that data were affected by low proficient students with less clear idea of the reading skills and strategies used, the researchers decided to consider the information of the top group, i.e., the most proficient students. The results, as expected, were more consonant with their expectations.

The introspective study, based on the verbal reporting and thinking aloud, provided relevant information concerning: a) typical examples of the expected performance of using a particular skill/strategy; b) examples of unexpected performance of text processing and task completion; c) the general impression of the student's use of background knowledge, language competence and use of skills/strategies; and d) text processing and task completion performance, including reading style (selectively, expeditiously or carefully), contributory reading monitoring skills or strategies used, item responses (whether the item was correctly answered), and process of arriving at a correct or wrong answer.

Based on the study, the researchers concluded that the typical performance of reading shows that there are separable and different skills and strategies employed for the different reading purposes, which is clear in the case of careful global, careful local, and expeditious local reading, but not clear in search reading and skimming for top and bottom groups, either because of the above-average or below-average linguistic proficiency. The intended skills and strategies were not, thus, necessarily used by either top groups, since they always outperformed what was expected, or bottom groups, since they failed using the skills or strategies, but matched the performance of the middle group.

In sum, the researchers concluded that there is evidence to support a componential view of reading, and also to support some correspondence of method/item used and skill/strategy assessed, particularly high for the group with middle level of proficiency. This evidence is relevant for this research since it allows this researcher to analyze the performance of the items based on some correspondence of item used skill assessed, using the constructs as defined in this chapter, and the method of argumentation, as discussed in chapter 4.

CHAPTER IV

The Study

In the present study, I aim at investigating the defensibility of items used to assess reading ability in English as a foreign language, included in tests administered in three different testing situations: proficiency tests, university entrance examinations, and EAP classroom tests.

For that investigation, in this chapter I present the method to be used in this research for the analysis of the test content, more specifically test items. The analysis will be based on the approach to validation proposed earlier as logical validity (Cronbach, 1949, as cited in Messick, 1989), plausible rival hypotheses (Campbell, 1957, as cited in Shepard, 1993) and as the notion of falsification (Popper, 1962, as cited in Shepard, 1993), and later revisited by many scholars including Messick (1989), and Shepard (1993) with the idea of plausible competing explanation, and also elaborated as argument-based approach by Kane (1992), and as validity argument by Chapelle (1994, 1999). In the cases where the notion of validity was revisited, validation requires justification of the interpretation and use of the test score, and this justification should include evidence that the test score reflects the area(s) of language ability to be measured, and nothing else, as well as considerations of the consequences of test use, both intended and unintended.

First I define test validation and describe sources of validity evidence as put forth by Messick (1989) (section 4.1). Then, I discuss the recent method of validation as argument-based, which has been suggested to support the validity of the interpretations and uses (section 4.2). In section 4.3, I present an articulation of method and source of

evidence. In section 4.4, I delineate the sources of evidence used in this research. In section 4.5, I specify the focus of the analysis. In section 4.6, I define the criterion for criterion-related evidence, and, in section 4.7, I describe and define the facets of the framework by Bachman (1990) and Bachman and Palmer (1996) to assess tasks characteristics to be used for the analysis in this research. In section 4.8, I present the approach of Reverse Engineering by Davidson and Lynch (2002) as an approach for the validation process, making it clear that the direction of my analysis will be from copies of tests to the supposed construct assessed through each test item. In section 4.9, I display the material to be used for the analysis.

4.1 – Validation and Source of Validity Evidence

Test validation, according to Messick (1989) is a “process of *inquiry* into the adequacy and appropriateness of interpretations and actions based on test scores” (Messick, 1989, p. 31). The process of validation implies necessarily collecting evidence of validity, i.e., finding sources of evidence, and also implies providing justifications for the interpretations and actions based on the test scores.

According to Messick (1989), the process of validation is a continuing process, and involves procedures of making a case to provide support for the interpretive inferences and action inferences based on the test score. In case of the validation of an interpretive inference, multiple lines of evidence must be provided in support of the proposed inference and also of discounting the alternative inferences. In case of the action inferences, value implications and action outcomes, in terms of the relevance and utility of the test score, must be considered. Validation, thus, implies “the interpretability, relevance, and utility of scores, the import or value implications of

scores as basis for action, and the functional worth of scores in terms of social consequences of their use” (p. 13)²².

Still concerning validation, Messick claims that it is a “scientific inquiry into score meaning, that score-based inferences are hypotheses and the validation of such inferences is hypothesis testing” (p. 64). These hypotheses, he asserts, are originated from two different sources: construct theories of performance domains and construct theories of the critical aspects of performance, referring respectively to the domain as content of a course, and to the domain as criterion of expected performance in the future situation of performance.

In any case, according to Messick (1989), it is the construct theories, as domain construct or criterion construct, in terms of the knowledge, skills, cognitive processes, or personal attributes implicated in successful performance, which will provide the basis to guide the hypothesis to be tested. This central and unifying role of construct meaning of test scores in all the treatment of validity is clear in his following statement:

since construct-related evidence undergirds not only construct-based inferences but content- and criterion- based inferences, as well as appraisals of testing consequences, construct interpretation is the unifying force that integrates all facets of test validity” (p. 89).

Collecting validity evidence is related to analyzing the content of a test and, more specifically, to examining carefully the meaning of performance obtained through the items chosen for a test. The ultimate purpose is to ensure the defensibility of the test content, and more specifically its items, for the interpretation and action based on the test taker’s performance. As McNamara (2000) puts it

the purpose of validation in language testing is to ensure the defensibility and fairness of interpretations based on test performance. It asks, ‘on what basis is it proposed that individuals be admitted or denied access to the criterion setting being sought? (p. 48).

²² Messick resorts to the etymology of the words *valid* and *value*, claiming that both have the same Latin root, *valere*, meaning ‘to be strong’, but that the derivation of the old French word *valoir*, meaning ‘to be worth’ is best applied to the current meaning for the word *valid* as referring to “the functional worth of the testing” (p. 59).

The part of the question posed by McNamara *on what basis* claims considerations as to the evidence collected to support the interpretation inferences and action inferences based on the test score.

Messick (1989) has contributed in terms of providing ways for collecting the relevant information or evidence needed for the validation process. He claims that the following sources can be used individually or in some combination as evidence for validation: a) comparison of the test content to the content of the domain of reference; b) probing of the ways responses are given to the items or tasks; c) examination of the internal structure of test responses, i.e., the relationships among responses to the tasks or items; d) survey of the test's external structure, i.e., the relationships of the test scores with other measures and background variables; e) investigation of the differences in response to experimental manipulations; and f) investigation of the social consequences, both intended and unintended, of interpreting and using the test scores in certain ways (p. 16).

In this research, I will focus mainly on the second source of validity evidence mentioned above, namely, probing the ways responses are given to items or tasks, but will also consider the first and the last sources mentioned, namely, the comparison of the test content to the content of the domain of reference, and the investigation of the social consequences, both intended and unintended, of interpreting and using the test scores in particular ways.

According to Messick, validation approach involves the following processes for the three sources of evidence: for the probing of the ways responses are given to items or tasks, the process involved is to directly probe the processes underlying item responses and task performance; for the comparison of the test content to the content of

the domain of reference, the process involved is to engage in judgmental and logical analysis as is done in documenting content relevance and representativeness; and for the investigation of the social consequences, both intended and unintended, of interpreting and using the test scores in particular ways, the process involved is to appraise the value implications and social consequences of interpreting and using the test scores in particular ways (p. 49).

4.2 – Methods of Test Validation: Arguments and Justifications

Generally, the method for test validation, that is, how to produce evidence, begins, according to Chapelle (1999), with a “hypothesis about the appropriateness of testing outcomes (inferences and uses)” (p. 259), and involves the collection of evidence related to the hypothesis, and the organization of the arguments for which a validity conclusion is drawn. Justification must be provided for the interpretation of test score by considering its construct validity (evidential for interpretation) and the value implication of interpreting this score in a particular way (consequential for interpretation). Justification must be provided for the use of score by presenting evidence, or arguing coherently, that the ability is essential to the individual’s performance (evidential for use), and by considering the consequence (intended and unintended) of the decisions made based on the test scores (consequential for use) (Bachman, 1990, p. 242).

The definition of validation as hypothesis testing of the hypotheses based on the meaning of the test score has resulted in approaches to validation, such as Kane’s (1992), and Chapelle’s (1994, 1999). They are explained below and their contribution to this research is made clear.

Kane's (1992) argument-based approach

Kane (1992) advocates the use of what he calls an argument-based approach to validation and validity, involving interpretive arguments. The idea to use the term *argument* lies in the core of his proposal, since it implies, according to the author, persuading an audience with a positive case for the proposed interpretation, and also implies considering and evaluating competing interpretation (p. 534).

Accepting the more recent idea that validity is associated with the interpretation of the test scores rather than with the test itself or its scores, the author makes a point in defining interpretation as involving meaning or explanation, or rather, explanation of meaning, involving an argument leading from the scores to score-based statements or decisions. Validity, in this case, would depend on the plausibility of the arguments. As the author puts it,

To validate a test-score interpretation is to support the plausibility of the corresponding interpretive argument with appropriate evidence. The argument-based approach to validation adopts the interpretive argument as the framework for collecting and presenting validity evidence and seeks to provide convincing evidence for its inferences and assumptions... (p. 527)

As it is possible to see, the two terms – argument and interpretation – become the key concepts within his proposal, which he calls interpretive argument.

Kane (1992) explains further the processes leading to validity and validation of the statements and decisions based on the test scores as: the first step is to outline the statements and decisions based on the test scores; the second is to determine the inferences and assumptions underlying the statements and decisions; the third is to identify potential competing interpretations; and finally, the fourth is to find evidence

as support of the inferences and assumptions in the interpretive argument and to refute potential counterarguments. In general lines, the procedure for the argument-based approach involves: “one chooses the interpretation, specifies the interpretive argument associated with the interpretation, identifies competing interpretations, and develops evidence to support the intended interpretation and to refute the competing interpretations” (p. 534).

Kane (1992) compares the kind of reasoning underlying practical argument he advocates to the kind of reasoning used for traditional logic and mathematics. As to the latter kind of reasoning, he stresses that “the assumptions are taken as given, and the conclusions are proven (i.e., the proof is *logically valid*), if and only if the chain of inferences from the premises to the conclusions follow certain explicit, formal rules” (p. 528), whereas the former kind of reasoning further involves inferences and assumptions evaluated with supporting evidence concerning the “appropriateness of various lines of argument in specific contexts, the plausibility of assumptions, and the impact of the weak assumptions on the overall plausibility of the argument” (p. 528).

Three criteria are used for the evaluation of practical arguments: 1) clarity of the argument – requires that the inferences and assumptions be specified in detail so that the argument is clear; 2) coherence of the argument – requires that the conclusions follow from a reasonably specified assumption; and 3) plausibility of assumptions – requires that the argument be inherently plausible or supported by evidence.

Concerning the relevance for validity, Kane (1992) maintains that interpretive arguments have four general characteristics. They: 1) are artifacts; 2) are dynamic; 3) may need to be adjusted; and 4) are practical argument evaluated in terms of degree of plausibility. The first characteristic refers to the fact that different interpretations may be needed depending on the variation along the several dimensions that can possibly be

used for the interpretation. The second refers to the fact that as more evidence is available, the interpretative argument may expand to include new types of inference, or may become smaller, if assumptions are refuted.

As to the third characteristic, the author claims that it should reflect the fact that test takers and situations may be different, since many assumptions are made based on what is plausible under normal circumstances for ordinary test takers. The fourth characteristic should reflect what distinguishes interpretive argument from logic or mathematics, i.e., it is necessarily judgmental, and the conclusions involve the plausibility of the argument rather than a *yes* or *no* answer or decision. As the author summarizes it, “interpretive arguments are artifacts, they change with time, they may need to be modified for particular examinees or circumstances, and they are more-or-less plausible” (p. 533).

Kane (1992) emphasizes that there are many advantages to this approach in comparison to other approaches for validation. First, an interpretive argument may be associated with formal theory but not necessarily; second, it is highly tolerant, hence involving any kind of interpretation or use of data collection techniques; third, the plausibility of the interpretive argument can be improved with more evidence supporting or not the most questionable inferences and assumptions; and fourth, the measurement procedure can be improved since the approach focuses on specific parts or aspects of the procedure.

Chapelle's (1994) validity table - validation as justification

Chapelle (1994) has become one of the most influential advocates of the approach of validation as justification. She draws on the debate on validity and

validation, in particular on Messick (1989), and eventually develops what has come to be known as the validity table, which is presented below as table 5

Table 5: Validity table presented by Chapelle (1994)

Justifications	Argues in favour	Argues against
Evidence		
Consequences		

The author places justification as a central element for validity inquiry, and, in consonance with Messick's (1989) definition of validity as the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on the test scores, she defines justification as providing empirical evidence and theoretical rationale based on the construct used in the test.

As to the evidence for construct validity, Chapelle says that it refers to judgmental and empirical justifications supporting the inferences (interpretation inferences as well as action inferences) made from test score, that is, construct validity evidence must be examined to assess the extent to which the examined evidence demonstrates that components of the ability are responsible for performance in the test.

It is possible, thus, to say that the evidence to be collected must have the following basic assumptions: a) if items are answered correctly because of some aspect of the ability being assessed (reading ability), they are measures of the ability; and b) if items are answered incorrectly because of the lack of some aspects of the ability being assessed (reading ability), they are measures of that aspect of the ability. These assumptions must orient the arguments in support of the score-based interpretations.

Adding to Chapelle's Validity Table

The process of validation must be seen as an ongoing process of collecting evidence *for* and/or *against* the inferences based on the test score, which, in turn, is a reflection of the performance elicited through the methods and/or items in a test.

Keeping this notion of an ongoing process, Davidson (personal communication, March, 2004) suggests the addition of a third column in the validity table, so that it is possible to enrich the discussion during the validation process. The three-column table is presented as table 6 below:

Table 6: Validity table with the extra column of the refutation of the argument against

Justifications	Argues in favour	Argues against	Refutation of the argument against

In my analysis, I will use this three-column table for the argument-based validation, and may add comments as the *refutation of the argument against* whenever relevant and possible. The plausibility of the arguments will determine the degree of the validity of the interpretation based on the performance elicited through the item.

4.3 – Articulating Method to Sources of Evidence

Based on Messick's (1989) idea that inferences and the validation of inferences is hypothesis testing, and on his six basic sources of validity evidence, Chapelle (1999) proposes that the process of validation must go through the following three steps:

making hypotheses about the testing outcomes, defining the source for collecting the relevant evidence for testing the hypotheses, and developing a validity argument.

The first step refers to hypotheses making about testing outcomes, more specifically, what the test is expected to test, and what their scores mean, i.e., hypotheses related to the assumptions about what a test measures and what their scores can be used for. This requires the specification of the construct underlying the test, since hypotheses about performance on a test are derived from a theory of what the task of responding to a question involves. Thus, the specification of a construct theory is essential, since it is based on it that hypotheses can be developed and it is against it that evidence will be evaluated.

The second step, defining the source for collecting the relevant evidence for testing the hypotheses, is mostly based on Messick's (1989) six basic sources of validity evidence. Her explanation²³ of how each of them could become approaches to collect the relevant construct evidence for the validation process for language testing is presented below for the three sources of construct evidence to be used in this research.

Considering the source of evidence as the probing of the ways responses are given to items or tasks, her approach implies relying on empirical analysis of the responses, not judgmental analysis, and assessing if the hypothesized knowledge and processes are responsible for the performance. The analysis may be quantitative or qualitative. The quantitative is based on both item difficulty and discrimination value. The qualitative analysis is based on the hypothesis that the test taker is really engaging in construct-relevant processes while taking the test.

Considering the source of evidence as the comparison of the test content to the content of the domain of reference, Chapelle's (1999) approach involves experts'

²³ This explanation is also presented by Chapelle (1994).

judgment of the content of the test in terms of its relevance, representativeness, and technical quality, i.e., it involves content analysis to provide “evidence for the hypothesized match between test items or tasks and the construct that the test is intended to measure” (p. 260), construct which, according to the author, must be explicit to guide the analysis.

Considering the source of evidence as the investigation of the social consequences, both intended and unintended, of interpreting and using the test scores in particular ways, her approach implies assuming that testing consequences have value implications, and involve hypotheses about how the test impacts all the people involved with it, thus going beyond test inferences. Any of the sources of evidence in any combination may be collected, allowing for the continuation to the third step in the process of validation.

The third step in the process of validation is the development of a validity argument. A validity argument must “present and integrate evidence and rationales from which a validity conclusion can be drawn pertaining to particular score-based inferences and uses of a test” (Chapelle, 1999, p. 263). She stresses that a validity conclusion is an argument-based and context-specific judgment, although she recognizes the challenges to the method, in particular the problems in identifying the appropriate types and number of justifications, and the problems to integrate them in drawing a validity conclusion, as well as the fact that test use is context-specific.

4.4 – Defining the Sources of Evidence

In this research, I will be focusing mainly on the probing of the ways responses are given to items or tasks, but will also consider the comparison of the test content to

the content of the domain of reference, and consider the social consequences, both intended and unintended, of interpreting and using the test scores in particular ways. In combination, these steps will provide the evidence needed for the validation process, and arguments and justifications will be given and considered for a validity conclusion to be drawn.

Comparison of the test content to the content of the domain of reference

For the source of evidence coming from a comparison of the test content to the content of the domain of reference, I will conduct an analysis considering the test content in terms of skills and abilities included in the test as compared to the domain of reference of the construct of language ability and the construct of reading ability as discussed in chapter 3, with the use of the framework for analysis of task characteristics proposed by Bachman (1990) and Bachman and Palmer (1996), mentioned in chapter 2 and recapitulated below in this chapter.

Focusing on the construct is justified within the most recent perspective for validity, where it is seen as a unified approach, with construct as the unifying element. However, both content- and criterion- related evidence may be analyzed. This does not go against the unified approach, because the content of a course is, or should be, in fact, the construct delineated for a particular course, and the criterion is, or should be, the construct delineated as the essential behavior for a particular use of language in real life.

Thus, the domain of reference may be more specific as the content of a course or the criterion determined as essential behavior for future language use, and more general, in terms of the construct as determined by theories of language ability and

reading ability. However, since content and criterion are delineated by a construct, all the analysis will be based on the constructs used in this research.

Considerations of the social consequences of interpretation and use of the test scores

For the appraisal of the social consequences, both intended and unintended, of interpreting and using the test scores in particular ways, I will consider Bachman's (1990) notion of test bias, Bachman and Palmer's (1996) notion of fair testing, and Shohamy's (2001) critical testing approach, all discussed in chapter 2.

Probing of the ways responses are given to items or tasks

For the source of evidence coming from the probing of the ways responses are given to items or tasks, my main focus, I will consider myself as an expert probing answers, and analyzing the processes underlying that. It will be a task analysis, empirical, taking into consideration the research on the correspondence of item/method used – skill/ability measured, discussed in chapter 3.

Process analysis is considered by Messick (1989) a possibility for gathering evidential basis for the validation process, because of the several techniques for the direct analysis already available, such as the protocol analysis, chronometric analysis, cognitive correlates of test performance, analysis of reasons, analysis of eye movements, and analysis of systematic errors (p. 53).

Looking specifically at the field of language testing, Bachman (1990) agrees that process analysis, in terms of protocol analysis, analysis of reasons for choosing a particular answer, and analysis of systematic errors, allows the investigation of the processes involved in the performance of tasks. However, Bachman adds another

qualitative method called self-report data, and claims that the studies in which the method was used “clearly demonstrate its usefulness for permitting us to better understand what the test takers actually do when they take the tests, and hence, what it is that our language tests actually measure” (p. 270).

This procedure is similar to what Alderson et al (1995) have called Response Validity, defined as the information on how test takers respond to the items in the test, “the processes they go through, the reasoning they engage in when responding” (p. 176). The method for collecting this information, according to them, is introspection, which should be collected retrospectively, by using interviews about the reasons why test takers have produced their answers.

Talking specifically about an example of use of the test method called cloze task, Alderson et al add that introspection shows “whether the student has to answer an item by using the range of reading skills intended by the test designer, or whether all that is needed is some knowledge of the grammatical structure of the phrase in which the item appears” (p. 176).

Using the method of introspection would, according to the authors, reveal the problems with test items, since some items may produce wrong answers in spite of the test taker’s understanding of the passage, or conversely, some items may produce right answers in spite of the test taker’s misunderstanding of the passage (p. 176). Based on that, it is possible to say that this method of introspection may be part of the validity argument, response validity as called by the authors.

I recognize two limitations for collecting validity evidence using the aforementioned process analysis, self-report data, and/or introspection, all including analysis of reasons for answers. The first limitation refers to the fact that expert judgment should not be considered the only evidence for validation, which is in

agreement to Messick's (1989) claim that the investigation of whether or not test items tap relevant knowledge or skills "cannot be left to the supposition or expert judgment alone" (p. 70). However, expert judgment of item relevance and representativeness, as Messick himself claims, is only a starting point of the test evaluation (p. 70), thus of validity investigation. It is this starting point that I aim at with my analysis of item performance within this validation task.

The second limitation is that different test takers may use different skills and/or strategies to solve the same tasks as discussed in chapter 3. My analysis will be mostly interpretive, considering the arguments in favor and against the inferences likely to be drawn through the use of test items. It will take into consideration the alleged nature of task performance in test, but also Weir et al's (2000) findings that there is a consensus among experts of what each test item is likely to measure, and that there is a coincidence of this consensual agreement with middle-level test takers' performance in tests as shown by the introspective and retrospective analysis within their study, as discussed in chapter 3. And these middle-level test takers may be those most tests should be aimed at, since they are the test takers likely to have the minimum requirement for the reading tasks in the criterion or target language use situation.

This concern for the limitation of generalizing results is in line with what Bachman (1990) states to be part of the process of construct validation when he says that the process of construct validation will result in a "statement regarding the extent to which the test under consideration provides a valid basis for making inferences about the given ability with respect to the types of individuals and contexts that have provided the setting for the validation research" (p. 271).

Since I will be the individual, and the context is this research, the statements should reflect this, and caution must be exercised when generalizing this analysis to

other testing situations. However, I will analyze item performance and the inferences about ability/skills assessed, as researchers have tended to do more recently, probably recognizing the difficulty of the task, in particular after the publication of Alderson (1990a) and (1990b) discussed in chapter 3: when discussing the correspondence between test item and ability(ies)/skill(s) assessed, they use *items dealing with... may depend on...* (Weir 2003, p. 134-5), and *appear to measure* (Bachman et al, 1995 p. 24 and 82). I will follow the same procedure.

4.5 – Defining the Focus of the Analysis: Relevance and Representativeness

Construct validation inquiries about the adequacy of the content of a test. The adequacy of the construct present in the test will be investigated considering the aspects of relevance and representativeness, which can only be judged considering the domain of reference (Messick, 1989, p. 37).

Domain of reference, used as a guide for the adequate selection of items to be included in the test, has been defined as the “total body of information for which the construct is expected to account” (Messick, 1989, p. 37), and is delimited by the use of a particular test. The domain of reference may be more specific as the content of a course – the syllabus, or the criterion determined as essential behavior for future language use, and more general, in terms the construct as determined by the constructs of language ability and reading ability presented in chapter 3.

Messick (1989) stresses, however, that, although these situations have the specifics concerning content definition and criterion definition, they also have their similarities since, in both situations, the definitions of the content for a syllabus or of the criterion for future performance are, or should be, based on some construct. In both

cases, tests are, or should be, ultimately based on some construct, and the focus of validation should be the same as construct validation with the specifics. Thus, the analysis in this research will have the construct of reading ability as its basis.

The adequacy of the test content – relevance and representativeness – in relation to the construct will be judged, in this research, considering the two following aspects: tests may either not cover some relevant aspect of the construct, or may cover some irrelevant aspects, or both. As Messick (1989) points out,

tests are not only imprecise or fallible by virtue of random errors of measurement but also inevitably imperfect as exemplars of the construct they are purported to assess. Tests are imperfect measures of constructs because they either leave out something that should be included according to the construct theory or else include something that should be left out, or both. (p. 34).

Both leaving out relevant aspects and/or including irrelevant aspects have been considered threats to construct validity. When tests leave out some relevant aspect of the construct, or when the tests require little of the candidate, the threat is referred to as construct underrepresentation (McNamara, 2000, p. 53). When tests include something that should have been left out, or when tests introduce factors irrelevant to the aspect of ability assessed, the threat is referred to as construct irrelevance (McNamara, 2000, p. 53).

The two threats may be manifested in the following way: for the construct-irrelevant test variance, there might be, for example, construct-irrelevant difficulty and/or construct-irrelevant easiness; for construct underrepresentation, there is the multiple-measure versus mono-measure approach to the construct.

Construct-irrelevant difficulty refers to aspects in the tasks not relevant to the focal construct that “make the test irrelevantly more difficult for some individuals or groups” (Messick, 1989, p. 34), such as subject matter of the text or test format.

Construct-irrelevant easiness refers to the opposite, i.e. to aspects not relevant to the focal construct which make the tasks easier, allowing some individuals to “respond correctly in ways irrelevant to the construct being assessed” (Messick, 1989, p. 34), such as clues in the items or test format. Messick (1989) stresses that they are both threats to validity for the interpretation of score meaning: “...both construct-irrelevant difficulty and construct-irrelevant easiness, when they occur, are important sources of invalidity with respect to construct interpretation” (p. 35).

The idea of construct-irrelevant easiness is similar to Popham’s (1981) unintended clues, which are a problem for validity. As the author stresses,

an obstacle to the creation of stellar test items arises when writers inadvertently toss in clues which permit examinees to come up with the correct answers to items that they couldn’t answer correctly without those unintended clues. If there are many of these unintended clues in a test, the test’s validity will surely be impaired (p. 239).

Construct underrepresentation may be manifested by the use of one measurement, since each measure individually may be underrepresentative of the whole construct, whereas multiple measures might converge various sources of evidence into a composite, increasing the possibility of covering the whole construct.

In sum, the threat to the adequacy of test content in terms of its representativeness is construct underrepresentativeness or underrepresentation, i.e., the test is too narrow and fails to include important dimension or facets of the construct, and the typical question is *to what extent does test content adequately represent the domain of reference?* The threat to the adequacy of test content in terms of its relevance is construct irrelevance, i.e., the test contains variance that is irrelevant to the interpreted construct, and the typical question is *to what extent is the test content relevant to the domain of reference?*

Since relevance and representativeness must be judged considering a domain of reference, I consider, next, the focus of the validation process in relation to the types of testing which are analyzed in this research.

According to McNamara (2000), there are two main uses for tests: tests assessing achievement, called achievement tests, and tests assessing proficiency, called proficiency tests. Achievement tests are based on previous content, usually related to some course syllabus, thus being “retrospective, giving evidence on what has been achieved” (p. 49). Proficiency tests provide information for the inferences to be made about future performance in the criterion setting, thus being “predictive or forward looking” (McNamara, 2000, p. 49).

Considering the testing situations analyzed, it is possible to say that classroom tests are achievement tests, TOEFL and IELTS are proficiency tests. However, a third type must be added to McNamara’s twofold division: selection tests. This is the case of the university entrance examinations in Brazil, used for selection for fixed-quota places for undergraduate studies.

In all the three testing situations, the analysis of relevance and representativeness will consider the construct as dictated by the constructs of language ability and reading ability presented in previous chapters.

For the entrance examinations, the analysis of relevance and representativeness will also consider criterion-related evidence, since predicting future performance in the criterion is part of the purpose of university entrance examinations. The criterion considered is further explained below.

I will also consider, for the university entrance examinations, the degree of correspondence between the test tasks and the criterion tasks. For that, I will use Bachman’s (1990) framework for the analysis of the authenticity of the tasks, whose

degree determines the kind of generalization possible, i.e., high degree of correspondence means authenticity of test tasks in comparison to the target language use tasks, which allows the generalization from performance on the test to performance in the domain of reference, or from performance to competence. More on Bachman's framework is in section 4.7 below.

In addition to that, for all the three testing situations studied in this research, I will also consider the technical quality of the items as a general evaluation, and provide my expert judgment for features such as readability, freedom from ambiguity, appropriateness of keyed answers and distractions, and clarity of instructions (Messick, 1989, p. 39). Test items must be free of these technical quality problems for the validity to be considered, and they will be considered defensible if they allow performance to reflect the level of ability of test takers, not technical problems.

In sum, in this study, the construct of reference is language ability, and more specifically reading ability. Both relevance and representativeness will be judged based on the accounts expounded in chapter 3.

4.6 – Defining the Criterion for the Analysis of the University Entrance Examinations

Defining the criterion is essential for the validation process, in particular for the collection of criterion-related evidence. Criterion-related evidence requires the examination of the relationship between test score and some criterion indicating the ability being tested, the criterion being the performance in the target language use situation, i.e., reading for academic purposes.

Bachman (1990) points out that there is a potential problem for collecting criterion-related evidence: the indeterminacy of the future situation, resulting from the

inability to identify the abilities and factors relevant to the criterion, and from the inability to specify how they relate to each other. While I agree with Bachman as to the difficult task of determining the criterion specifications, I also agree with one possible solution presented by himself, i.e., the solution of simplification, requiring the reduction of the number of measures to small sets (p. 251). This is appropriate for this research, since it is restricted to the one skill of reading, rather than the whole of language ability as considered by Bachman.

For the task of defining the criterion for criterion-related evidence, the criterion being the performance in reading for academic purposes, I will use my own experience as a student, my experience as a teacher of EAP reading courses, and the contribution by Urquhart and Weir (1998), and Weir et al (2000) with their taxonomies of reading for academic purposes based on their study of the needs analysis of university students.

The characteristics of the criterion for criterion-related evidence are summarized and shown in table 7 below, for general conditions.

Table 7: Specification for university studies: general conditions for academic purposes based on Weir et al (2000)

Conditions	Descriptions
Purpose(s) of reading	To comprehend academic texts and to extract important information
Nature of texts	Informative texts ²⁴ .
Source of texts	Books, journal articles, abstracts, theses, dissertations
Rhetorical organization	Mainly expository texts with rhetorical organization of comparison, collection of description, problem/solution, and causation
Illocutionary features	To inform, to explain, to describe, and perhaps to advise
Channel of presentation	Normally textual, with possible graphics

The characteristics of the criterion for criterion-related evidence based on Urquhart and Weir (1998) are presented in table 8 below for the skills and strategies specific for both reading types – expeditious reading and careful reading.

²⁴ Davies (1995) classifies encyclopedias, textbooks, academic papers, specialist journals as informative in her classification of genres with reference to primary social function and reader purpose (p. 130).

Table 8: Specification for university studies: skills and strategies for each type of reading (Urquhart & Weir, 1998).

Type of reading	Skills and strategies
Expeditious reading	<p>Skimming quickly to establish discourse topic and main ideas.</p> <p>Search reading to locate quickly and understand information relevant to the predetermined needs</p> <p>Scanning to locate specific information; symbol or group of symbols; names, dates, figures or words</p>
Careful reading	<p>Reading carefully to establish accurate comprehension of the explicitly stated main ideas the author wishes to convey</p> <p>Understanding syntactic structure of sentence and clause.</p> <p>Understanding lexical and/or grammatical cohesion.</p> <p>Understanding lexis/deducing meaning of lexical items from morphology and context</p>

The characteristics of the criterion for criterion-related evidence based on Urquhart and Weir (1998) are presented in table 9 below for the reading purposes specific for the skills and strategies.

Table 9: Specification for university studies: purposes for skills and strategies for each type of reading (Urquhart & Weir, 1998).

Skills/strategies	Purposes
Skimming	<p>Establish a general meaning</p> <p>Establish the macropropositional structure</p>
Search reading	Locate relevant information for careful reading
Scanning	Locate specific information
Reading carefully for explicitly stated main ideas	<p>Establish the macrostructure</p> <p>Reading to understand it all and learn</p>
Reading carefully for implicitly stated main ideas: inferring propositional meanings	<p>Discover writer's intention</p> <p>Understand writer's attitude to the topic</p> <p>Identify the addressee</p> <p>Distinguish fact from fiction</p>
Reading carefully for meaning related to the text: inferring pragmatic meanings	<p>Apply main ideas to other contexts</p> <p>Evaluate a point of view</p> <p>Express own opinion on the subject</p>

There are, however, two problems with these definitions for the criterion. One refers to the fact that it may be difficult to define specific purposes for all university students. Alderson (2000) has claimed that it is impossible to talk about one specific purpose and amount of reading for the various areas in the criterion of university studies, since mathematics or computing, for example, require little normal reading, engineering, chemistry and biology require only limited amount of reading, but linguistics, philosophy, literary studies, and history may only require reading in depth.

While I tend to agree with the author as to the amount of reading, I tend to disagree with him as to the purpose. I understand that the main purpose for all the university students to carry out their studies is study reading, reading with the purpose of learning the content or procedures, which requires “slower reading, reading in depth, and time for reflection” (Davies, 1995, p. 134).

The other problem refers to the finding, by Urquhart and Weir (1998), of the task of inferring pragmatic meanings as shown in table 9 above. This task in the criterion can be accounted for since university studies require the application of main ideas to other contexts, the evaluation of a point of view, the expression of opinions on various subjects. As discussed in chapter 3, this task of inferring pragmatic meanings requires background knowledge to an extent that traditional testing cannot incorporate, therefore they will be considered, in this research, as sources of invalidity. Norton and Stein (1998) refer to this paradox as a ‘validity paradox’, since testing instruments may not give test takers the opportunity to demonstrate abilities required in the criterion.

4.7 – Defining the Facets for the Analysis of Task Characteristics

The validation inquiry in this research will be assisted by the analysis of the authenticity of the tasks. To Bachman and Palmer (1996), as already discussed in chapter 2, authenticity is defined as “the degree of correspondence of the characteristics of a given language test task to the features of a TLU task” (p. 23).

Authenticity allows the investigation of the extent to which the interpretations based on the test performance can be generalized to situations other than the test itself. Higher degree of correspondence between the test task characteristics and the TLU task should be expected in a test for any investigation of validation, in particular for construct validation, since, ultimately, it is the construct which will give interpretation of competence to the performance on the test.

Authenticity will be investigated through the framework proposed by Bachman (1990) and Bachman and Palmer (1996), and the correspondence will be judged in terms of low, medium, or high. Since it is a very comprehensive framework, I will be looking only into some of the facets I consider relevant for the discussion of the correspondence between task in language use situation and task in testing situation, relevance which is discussed below.

Also, I will be using it only for the analysis of the university entrance examinations. I will not be using it for the analysis of the proficiency examinations – TOEFL and IELTS, since similar analyses have already been carried out for them by scholars such as Alderson (2000) and Douglas (2000). I will not be using it for the analysis of the classroom tests since they are syllabus-oriented tests.

I provide, in table 10 below, a summary of the task characteristics, a brief explanation as to what characteristics will be relevant for the analysis, and the

characteristics of the criterion for considerations of high authenticity for the analysis of the correspondence between the characteristics of TLU situation tasks and test tasks.

Further comments, when necessary, follow the table below.

Table 10: Framework for the analysis of the correspondence between the characteristics of TLU situation tasks and test tasks – tasks to be considered.

Characteristics of the test tasks		
1. Characteristics of the setting	Comprise the physical circumstances	
<i>1.1 Physical characteristics</i>	Place, seating, lighting, noise level, etc. Although TLU situation and testing situation will always be different as to this characteristic, I want to consider this as relevant for analysis in terms of the affective schemata.	To be considered
<i>1.2 Participants</i>	People involved in the task. Although TLU situation and testing situation will always be different as to this characteristic, I want to consider this as relevant for analysis in terms of the loading of affective schemata.	To be considered
<i>1.3 Time of task</i>	Time chosen for the task. Although TLU situation and testing situation will always be different as to this characteristic, I want to consider this as relevant for analysis in terms of the loading of affective schemata.	To be considered

To continue...

Table 10: Framework for the analysis of the correspondence between the characteristics of TLU situation tasks and test tasks – tasks to be considered. (cont.)

Characteristics of the test tasks		
2. Characteristics of the test rubrics	These are the characteristics which have relatively little correspondence between language use tasks and test tasks (Bachman & Palmer, 1996, p. 50).	
<i>2.1 Instructions</i>	Instruction must be well understood for the best performance, specially in case the test taker may not be familiar with the task types	
Language	Native, target.	Not used
Channel	Instructions in reading tests will always be written	Not used
Specifications of procedures and tasks	The clarity and explicitness of the instructions will allow test taker to perform at the level of ability. This is already incorporated in the technical quality table, becoming irrelevant in this one.	Not used
<i>2.2 Structure</i>		
Number of parts/tasks	Number of parts or tasks in the test	Not used
Saliency of parts/tasks	The extent to which the different parts are distinguished from one another. Considered irrelevant by the researcher.	Not used
Sequence of parts/tasks	The order of the parts. Considered irrelevant by the researcher.	Not used
Relative importance of parts/tasks	How parts differ in importance. Considered irrelevant by the researcher.	Not used
Number of tasks/items per part	The number of tasks included in each part. Considered irrelevant by the researcher.	Not used
<i>2.3. Time allotment</i>	Although important for the completion of the tasks, this researcher does not have the information for this characteristic to be used	Not used
<i>2.4 Scoring method</i>		
Criteria for correctness	When information is available, the researcher may use it for investigation of what kind of response is considered correct and, more importantly, why.	May be used
Procedures for scoring the response	When available, the researcher may use this information	May be used
Explicitness of criteria and procedures	When available, the researcher may use this information	May be used

To continue...

Table 10: Framework for the analysis of the correspondence between the characteristics of TLU situation tasks and test tasks – tasks to be considered. (cont.)

Characteristics of the test tasks		
3. Characteristics of the input	Input is all the material the test taker is expected to process. Thus, in this research, input may refer to both the texts and the tasks presented to the test taker	
<i>3.1 Format</i>		
Channel	Aural, visual It is always visual, thus irrelevant.	Not used
Form	Language, non-language or both. High: written texts with possible presence of illustrations as accessory to comprehension	To be used
Language	Native, target, or both. High: Target	To be used
Length	The length of input may influence the amount of required interpretation. The input may be a word, sentence, paragraph, or extended discourse. High: long extended discourse	To be used
Type of input	Item or prompt. Type will refer to the task only. High: item and prompt in the native language.	To be used
<i>3.2 Language of input</i>		
a) Language characteristics		
Organizational characteristics: Grammatical	Vocabulary, syntax, graphology High: vocabulary and syntax as determined in the academic writing	To be used
Organizational characteristics: Textual	Cohesion, rhetorical organization High: extensive use of cohesion devices for explicitness High: comparison/contrast, description, cause/effect	To be used
Pragmatic characteristics: Functional	Ideational, manipulative, heuristic, imaginative. High: ideational function	To be used
Pragmatic characteristics: Sociolinguistic	Dialect/variety, register, naturalness, cultural references and figurative language High: formal register, little cultural reference and figurative language	To be used
b) Topical characteristics	Personal, cultural, academic, technical. High: academic and technical	To be used

To continue...

Table 10: Framework for the analysis of the correspondence between the characteristics of TLU situation tasks and test tasks – tasks to be considered. (cont.)

Characteristics of the test tasks		
4. Characteristics of the expected response	Language use of physical response to the instruction, tasks, and input provided	
<i>4.1 Format</i>		
Channel	It is always visual, thus irrelevant.	Not used
Form	Language, non-language, or both High: language	To be used
Type	Selected response, limited production response, extended production response. High: limited response and extended production response	To be used
Language	Native, target, both High: native	To be used
<i>4.2 Language of expected response</i>		
	<p>The language of expected response is related to what the test taker is expected to show in terms of use of the language characteristics.</p> <p>In case of the limited or extended production, since this research is concerned with the receptive skill of reading in English as a foreign language, any assessment of productive skill in Portuguese or in English would be irrelevant.</p> <p>In case of the selected response given in English, the understanding of the language can be considered part of the input, and in case it is given in Portuguese, it is irrelevant.</p> <p>Thus, the characteristics under <i>language of expected response</i> are considered irrelevant for the present research</p>	
a) Language characteristics		Irrelevant
b) Topical characteristics		Irrelevant

To continue...

Table 10: Framework for the analysis of the correspondence between the characteristics of TLU situation tasks and test tasks – tasks to be considered. (cont.)

Characteristics of the test tasks		
5. Relationship between input and response.		
<i>5.1 Reactivity</i>	<p>The extent to which the input or response directly affects subsequent input or response</p> <p>Reciprocal - the responses change the form of the subsequent material</p> <p>Non-reciprocal - the responses do not change the form of the subsequent material</p> <p>Reading is always non-reciprocal (except for the instantaneous messages as used today)</p>	Irrelevant
<i>5.2 Scope of relationship</i>	<p>The amount or range of input to be processed for the response</p> <p>Broad scope – involving a great deal of input to answer, such as main idea comprehension questions</p> <p>Narrow scope – involving little amount of input to answer questions such as scanning, or lexical inferencing.</p> <p>High: both broad and narrow scope</p>	To be used
<i>5.3 Directness of relationship</i>	<p>The degree to which the expected response is based mainly on the information in the input or in test taker's own knowledge:</p> <p>Direct – response includes primarily information supplied in the input</p> <p>Indirect – response includes information not supplied in the input, for example, for opinions.</p> <p>High: both direct and indirect</p>	To be used

As to the facet *characteristics of the setting*, number 1 in table 10, including physical characteristics, participants, time of the task, it is possible to argue that there is low correspondence when comparing the test task characteristics and the TLU situation task characteristics, since the setting for the testing differs greatly from the target language use situation (normal reading situation).

The testing situation for the entrance examinations is very ritualized, with scheduled time known long in advance, a great deal of preparation, and pre-defined places. In fact, the higher the stakes, the more ritualized the testing situation is. Entrance examinations are very high-stakes tests.

As to the role played by the participants, in the testing situations, the participants are watchers representing a threat to the test takers. The participants in the case of reading in non-testing situation are usually non-existent, since reading for studying is usually silent (aside from group work).

Time of the task makes a difference since, in a situation of reading, it is the reader who chooses the best time for reading according to his/her personal characteristics and degree of tiredness. In a testing situation, the test taker has no choice other than complete the tasks as required at the time established.

It is, thus, possible to claim that there will always be low correspondence between the testing situation and the target language use situation in the testing situations analyzed for the facet of characteristics of setting. This low correspondence with all this ritualized testing situation often causes a great deal of stress to the test takers in both high- and low- stakes testing, but much more in high-stakes testing such as the university entrance examinations, which contributes to the loading of the affective schemata (with memories of tension) with negative influence on performance.

Considering that affect influences performance, as shown in the framework by Bachman and Palmer (1996) presented in chapter 3, and that, in the case of university entrance examinations analyzed, there will be a low degree of correspondence between the test task and TLU situation, there will always be a problem for validity, since unequal performance caused by unequal conditions will result in misleading evidence

for interpretation of ability or prediction of future performance, i.e., for generalization from the testing situation to the TLU situation.

Thus, although relevant for my considerations of validity, these characteristics become irrelevant for further analysis. They will, then, be considered as having low degree of correspondence with the consequence of low-authenticity testing situations in these facets, thereby compromising the generalization efforts from performance to competence or ability, i.e., limiting the possibility of using these low-authenticity testing situations to make interpretation inferences and action inferences based on the performance on the test.

The facet *characteristics of the test rubrics*, number 2 in table 10, have relatively little correspondence between language use tasks and test tasks, fact recognized by Bachman and Palmer (1996, p. 50). They are not, thus, to be used for authenticity analysis, and are not included in our framework for analysis of authenticity.

As to the facet *characteristics of the input*, number 3 in table 10, the correspondence in the testing situations should be determined by comparing the input (text and task) in the test to the input (text and task) the test takers will face in their academic life, that is, in the criterion or in the target language use situation defined above. In the case of the entrance examinations for the Brazilian universities, the input students will have are mainly written texts to be read in English for their academic studies.

Form of the input (language, non-language, both) refers to what form(s) of language are to be used in the input. This might include the written text as language form, and/or illustrations in general, such as pictures, diagrams, etc, as non-language form. The analysis in this research will focus on the assessment of language form, i.e., written texts, since this is the type of input test takers will mostly be using for their

future studies (in the criterion or TLU situation). The use of illustrations as part of the input is controversial. It can be argued that illustrations are part of some written texts, claiming for their use as authentic. It can also be argued that charts and tables, not pictures, are general features of expository texts (Weir et al, 2000). However, since the input in this research is language form – written texts, any tasks likely to focus on the use of illustrations only, or mainly, will be considered low authenticity, hence, a source of invalid evidence.

The language of the input in terms of the texts provided should be the target, in the case of this research, English, since it is the language to be used in the criterion, thus, part of the construct being measured in the testing situations analyzed. The language of the input in terms of the task might be the native, the target, or both. Since university professors in Brazilian universities are likely to give the assignments in the native language – Portuguese, with the TLU task in Portuguese, high correspondence would require the university entrance examinations to have their tasks also in Portuguese.

In this research, however, I would like to argue for the use of the target language – English – in the task (item or prompt) rather than the native, with the risk of reducing the correspondence, for the following reasons: using the native language – Portuguese – in the input might provide the test taker with a great deal of information in the native language related to the content of the text, which, in combination with other items or input such as illustrations, might allow the test taker to respond correctly to the item without reference to the text, the item, thus, becoming a passage-independent item. Using the target language in the task will demand reading the item (the task) in the target language, with reading ability in the target language being assessed, this way

avoiding the presence of construct-irrelevant easiness (Messick, 1989) or unintended clues (Popham, 1981), which are sources of invalidity evidence.

This is an example showing the need to consider each testing situation as determining the balance among the qualities of a test. In this case, it is this researcher's opinion that the relative weight of authenticity as assessed through this framework must be smaller than the weight of validity, authenticity, then, giving way to construct validity.

Considering the relative weight of each quality for each specific testing situation is in agreement with the proponents of the framework. Bachman and Palmer (1996) say that an optimum balance among the qualities must be achieved for each testing situation. In their words, "the relative importance of these different qualities [reliability, validity, authenticity, interactiveness, impact, and practicality] will vary from one testing situation to another, so that the usefulness can only be evaluated for specific testing situations" (p. 38).

The choice for the *type of input in terms of tasks* – item or prompt – should consider what has been referred to as a contaminating factor. Items require both selected or limited production, whereas prompt requires extended production. Limited production involves writing short answers, and extended production requires long answers, possibly essays (this relationship of task in terms of input and expected answer is further explained in the facet *characteristics of the expected responses* below). Although longer answers should prove more authentic if compared to the criterion demands, the construct in this research is reading ability. In this case, the longer the writing in the answer, the more authentic the task, the more invalid the evidence for the interpretation of reading ability. This is another example where authenticity should give way to validity.

As to the language of the input in terms of language characteristics, the analysis in this research will consider the grammatical characteristics of vocabulary and syntax, and textual characteristics of cohesion and rhetorical organization, as well as all the functional characteristics – ideational, manipulative, heuristic, and imaginative, and all the sociolinguistics characteristics – dialect/variety, register, natural or idiomatic expressions, cultural reference, and figurative language.

It can be said that academic expository texts have the following characteristics (also described in the section *defining the criterion* above): constrained by rules of good writing, formal vocabulary, good syntax, more presence of cohesion defining more explicitly the relationships between ideas and/or propositions, and common rhetorical patterns, described by Meyer (1975, as cited in Alderson, 2000), of collection, causation (cause and effect), response (problem-solution), comparison (comparison and contrast), and description (attribution).

Concerning language functions, although language use mostly involves the performance of multiple functions in connected discourse (Bachman & Palmer, 1996), it can be argued that, for the criterion or TLU situation, the language functions are mostly ideational, involving the expression and exchange of information about ideas and knowledge on all the subject matters in the academia, with descriptions, classifications, and explanations, but also with some instrumental function. It can also be argued that functions such as regulatory, interpersonal, and imaginative are mostly absent, although imaginative will be present in restrict areas, such as in literature. This is also considered in the section *defining the criterion* above.

As to the facet *characteristics of the expected response*, number 4 in table 10, there are some considerations to be made. In case of the language of the expected

response (native, target, both), the use of native language will provide evidence for authenticity, since, in the criterion, it is the native language that is used.

In the case of the item types, considerations related to the balance of the qualities as aforementioned must be made. While the use of limited response and extended production response is more authentic, since these responses are more likely to be required from university students, it imposes a problem for validity considerations for interpretations of reading ability, once they require producing language. Producing language through writing will be considered a contaminating factor for validity, since this research is on the assessment of the reading ability.

The same applies to the language of expected response, which is related to what the test taker is expected to show in terms of use of the language. In case of the limited or extended production, as this research is concerned with the receptive skill of reading in English as a foreign language, any assessment of productive skill in Portuguese or in English will be irrelevant. In case of the selected response given in English, the understanding of the language can be considered part of the input, and in case it is given in Portuguese, it is irrelevant. In sum, the characteristics under language of expected response are considered irrelevant for the present research.

As to the facet *relationship between input and response*, number 5 in table 10, the characteristic of reactivity is irrelevant, since, except for the instant messages in the chatting room on the Internet, reading has been non-reciprocal. Both broad and narrow scopes are relevant. The characteristic of directness of relationship is very relevant for the discussion, and it is related to the discussion presented previously in chapter 3, concerning the use of background knowledge.

Bachman and Palmer (1996) describe this facet as direct and indirect, the former including primarily information supplied in the input, and the latter, information not

supplied in the input, such as opinions. It is important to remember, however, that their framework is proposed for language ability, which involves an interaction of all the skills of listening, speaking, reading, and writing, or better, a non-separation perspective, for they do not endorse the separation of skills. In a communication event, all skills are involved. After reading, for example, the test taker would express orally his opinion, and this is considered language use. In the case of this research, reading is the only skill measured. Expressing opinions is beyond the scope of the construct of reading in that it requires the productive skill of speaking or writing, although it might be argued that it is authentic as a communicative event if compared to the criterion in which students are requested to participate with their opinions.

In addition to being beyond the scope of the construct of reading, opinion statement does not lend itself to being used in traditional testing because, in the end, the test user will have expected answers and will have to rate the test for correctness. This is the case for the university entrance examinations. Considering the balance among the qualities as discussed previously, authenticity must give way to construct validity, reliability (dependable information), and practicality required by large-scale tests.

4.8 – Reverse Engineering: an approach for collecting validity evidence

Validation concerns should be part of the whole process of the development of language tests. Considering that the validation process may have two different and opposite starting points, one from a construct in search of valid measurement and the other with a test in search of valid meaning (Messick, 1989, p. 89), in this research, since tests are the data, the starting point will be the tests, and the analysis will be in search of meaning.

Davidson and Lynch (2002) have elaborated on the approach for this procedure. They propose what they have called *Reverse Engineering*. The term engineering is used as an analogy to procedures adopted in civil or mechanical engineering. It is reverse since it refers to the opposite of the procedures adopted in the two fields. Thus, instead of producing the blueprint of a house or piece of equipment and arriving at the product, the opposite is done: from the existing product, the blueprint is produced.

In the case of language tests, this approach refers to the “creation of a test spec [specifications] from representative test items/tasks” (p. 41), where spec includes the skills or abilities tested. Thus, reverse engineering allows, based on the analysis of the performance of each item, for a comparison of the specifications and the abilities/skills intended with the specifications and abilities/skills really assessed, that is, a comparison between the real specifications with the supposed specifications, since there can be a difference between “the ‘real spec’ in use and the ‘supposed spec’ in an archive or testing manual” (p. 41). Reverse engineering is, thus, a process to arrive at the construct underlying the test based on the analysis of the operations and/or abilities underlying the performance elicited by each test item.

According to the authors, the approach allows to claim validity of a particular task: “the logic behind this approach to validation is that if the items or tasks for a test procedures are to result in valid inferences about test taker ability, then they should be readily identifiable in terms of the characteristics laid out in the test specifications” (p. 45), where the construct is identified.

In this research, I will be using this procedure of reverse engineering, since actual copies of tests are the data analyzed. Reverse engineering is the procedure adequate for the investigation of item performance when the researcher has only the actual copies of

the tests, without the responses from test takers, this being the adequate procedure for the present research (Davidson, personal communication, December, 2004).

The oriented experience I had previously with the procedure of reverse engineering includes the analysis of sample items following the procedure and the mode of argumentation suggested by Chapelle (1994) as the validity table, analysis which was presented, together with professor Davidson, to a group of students taking classes in language testing at the University of Illinois, USA, in May, 2004.

4.9 – Material used for data analysis

The proficiency tests analyzed are the TOEFL and IELTS. The analysis, in the case of the IELTS, is based on practice tests taken from a preparation book for the IELTS, published by Cambridge University Press. The analysis in the case of the TOEFL is based on practice tests taken from two preparation books published by Cambridge University Press, ETS, and Kaplan, and also on the copy of the pilot version of the test, made available on the Internet for analysis by ETS, to be released as the New TOEFL by September, 2005.

For the classroom tests, several tests were analyzed, which were used by three teachers teaching EAP (reading) classes at UFSC, SC, Brazil, and were chosen mainly due to their availability and willingness to provide the tests. This researcher understands that this method for choosing the teachers will not influence analysis, since no comparison or generalization is part of the objective of this research. Only the analysis of the tests provided by two of the three teachers will be reported here, one test by each teacher, chosen because of the diversity of their item types.

For the analysis of university entrance examinations, examinations, available on the Internet, of many universities were analyzed: 2002 and 2003 PUC-SP exams; 2000, 2001 and 2002 UFRJ exams; 2001, 2002 and 2003 UFMG exams; 2001 and 2002 USP exams; 1998, 1999, 2000, 2001 and 2002 UNICAMP exams; and 1998, 1999, 2000, 2001, 2002 and 2003 UFSC exams.

Examinations from only two universities were chosen for the analysis to be reported here: examinations of UFSC and UNICAMP, the former chosen due to the fact that it is done for the institution where I carry out my doctoral program, and the latter, because the examinations were introduced for “replacing the traditional multiple choice questions with open-ended questions” (Scaramucci, 2002, p. 64), and may be considered an innovative model for entrance examinations in Brazil. I will carry out the analysis of an entire examination of each university, and will also include item analysis from different examinations, in the case of UNICAMP, to illustrate some points within a validation study.

In the next chapter, I present the data analysis of the two proficiency tests and two classroom tests, considering the evidential basis as source of justification to judge the defensibility of the items.

CHAPTER V

Investigating the Defensibility of the Items in Proficiency Tests and EAP Tests:

Evidential Basis

In this chapter, I present the analysis of the proficiency tests and classroom tests. In section 5.1, I will present the analysis of the items within the IELTS tests, and in section 5.2, the analysis of the items within the TOEFL tests. Analysis of the classroom tests will be presented for Teacher 1, in section 5.3, and for Teacher 2, in section 5.4.

The items will be considered defensible if the evidence based on the performance elicited can support the inferences related to the interpretation of the reading ability according to the construct (interpretation inferences).

5.1 – Analysis of the IELTS test items

The IELTS is a battery of tests designed to assess the language proficiency of test takers, developed by University of Cambridge Local Examinations Syndicate (UCLES) as part of their English as a Foreign Language (EFL) tests (Weir & Milanovic, 2003, p. 64), now managed by UCLES, together with the British Council and the IDP Educational Australia Limited (Clapham, 1996).

The analysis of IELTS test items will be provided only of the academic reading module²⁵. The reading module is a multi-method test, having three texts followed by any one of the following methods: multiple-choice questions, short-answer questions, sentence completion, completion of tables/charts/summaries/notes, labeling a diagram,

²⁵ The IELTS examination has two different modules for reading assessment: academic reading and general training reading.

choosing headings for paragraphs or sections of the text, identification of a writer's view or attitude – yes, no, not given, classification, matching phrases/lists. Since the tests may use different methods, my analysis will comprehend all the methods, regardless of their being in one single test, and will not be necessarily based on one item²⁶. The analysis focuses on three tests – Test 1, Test 2, and Test 4 – reproduced in Appendices 1, 2, and 3 respectively. Next, I turn to the analysis of the items.

Method - Multiple-choice questions

Based on test 1, appendix 1, items²⁷ 5, 6, 15, 16, 17, and 40; on test 2, appendix 2, item 28.

Multiple-choice questions (MCQ), as used in this test, appear to assess any of the reading skills, stated or implied/inferred information, and global or detailed comprehension. Stated information takes the form of *according to the text*, or *according to the information in the text ...* as in test 1, items 6, 15, 16, and 17

Implied/inferred information takes the form of *which of the following statements best describe the writer's main purpose?* as in test 1, item 40, where the reader must infer the intention of the writer, and decide whether the writer intended to advise, encourage, explain, or help the students.

Global comprehension MCQ takes the form of *choosing the most suitable heading/title for a passage*, as in test 2, item 28, which requires integration and summarization of the information in the text, involving inferential processes.

²⁶ Since the analysis is not necessarily based on one single item, the items are not typed within the analysis, but they can easily be seen in the appendices 1, 2, and 3.

²⁷ They are called questions in the examinations

A detail-information MCQ will require reading for specific information, as in test 1, item 5, with the stem of *the greatest outcome of the discovery of the reaction principle was that*. A validity table for the method is presented below as table 11.

Table 11: Analysis of justifications for the method MCQ as used within IELTS

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to search for evidence of the ability to read for local and global comprehension	It allows guessing, compromising inference of ability	Distractors must be well developed to attract poor readers, thus increasing the discrimination index.

Is the item defensible? Yes. The method MCQ is flexible enough to allow the assessment of various skills, and different levels of comprehension, as shown in the analysis of the items above. If well developed and trialed, MCQ seems to have stronger arguments in favor of its use to provide evidence converging for the inference of reading ability. Further analysis of individual MCQ items is carried out below, within the analysis of the TOEFL test. Item 5, test 1, is further analyzed below.

Item 5, test 1

Item 5 is as follows:

Part of the text

...However, it wasn't until the discovery of the reaction principle, which was the key to space travel and so represents one of the great milestones in the history of scientific thought, that the rocket technology was able to develop. Not only did it solve a problem that had intrigued man for ages, but, more importantly, it literally opened the door to exploration of the universe.

The greatest outcome of the discovery of the reaction principle was that

- a) rockets could be propelled onto the air
- b) space travel became a reality
- c) a major problem had been solved
- d) bigger rockets were able to be built

A technical quality table for item 5 is presented below as table 12.

Table 12: Analysis of the technical quality of item 5, test 1, IELTS

Technical quality	Comments
Appropriateness of the key	<p>According to the book, the key is letter (b). However, it is too far from <i>it literally opened the door to exploration of the universe, or which was the key to space travel</i>. Both suggest that there would be changes in the future, and <i>became a reality</i> expresses a fact in the present.</p> <p>The idea that they are far events is also expressed more explicitly at the beginning of the following paragraph: <i>an intellectual breakthrough, brilliant though it may be, does not automatically ensure that the transition is made from theory to practice</i>.</p> <p>The last paragraph adds to this: <i>Nevertheless, the modern day space programs owe their success to the humble beginnings of those in previous centuries who developed the foundations of the reaction principle</i>.</p>

Is the item defensible? No. The item allows for low discrimination value, since it might confuse better readers, who eventually either choose the key by elimination, mostly like poor readers, or get the item wrong for the ‘right’ reason (having the ability). Items must allow performance (getting it right/wrong) to reflect ability. This item seems to have stronger arguments against its use to provide evidence converging for the inference of reading ability.

Method - Short-answer questions

Based on test 2, appendix 2, items 36-39.

Short-answer questions²⁸, as used in the test, require the test taker to find few words, one or two in the case, in the text. A validity table for the method is presented below as table 13.

Table 13: Analysis of justifications for the method short-answer question within IELTS

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to search for evidence of the ability to recognize vocabulary or to infer its meaning	It makes no reference to the information of the text.	Reader will need to read the text thoroughly to find the words.
		It requires writing.	It requires only copying words from the text, with the advantage that there are no clues for testwise strategies.
		It requires active use of vocabulary since it presents the context of meaning for the word to be provided.	The word to be provided is within its context in the text. The direction from meaning to word is the same for inference of word meaning.

Is the item defensible? Yes. Although reading is a receptive skill, not requiring productive use of words, this item allows for evidence of the ability to infer meaning of words. This item seems to have stronger arguments in favor of its use to provide evidence converging for the inference of reading ability, in this case, the ability to infer meaning of words.

²⁸ They can be considered what Pearson and Johnson (1978) called pseudo questions

Method - Sentence completion

Based on test 1, appendix 1, items 22-24.

Sentence completion item, as used in the test, will provide a number of incomplete sentences to be completed with phrases from a list, which includes distractors. A validity table for the method is presented below as table 14.

Table 14: Analysis of justifications for the method sentence completion within IELTS

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to look for evidence of the ability to read for specific information, to recognize propositions in the text and complete the sentence with their paraphrases.	Reading is not completing sentences.	Reading is relating information around a theme.

Is the item defensible? Yes. It assesses the ability to recognize propositions of the text and combine two parts of a sentence forming their paraphrase. Thus, it allows the assessment of grammatical knowledge, once the phrases to be chosen must relate syntactically with the last words given, as well as the assessment of the ability to follow the flow of argument, once the phrases must be related semantically. This item seems to have stronger arguments in favor of its use to provide evidence converging for the inference of reading ability, in this case, the ability to recognize propositions.

Method - Completion of tables, charts, and notes

Based on test 2, appendix 2, items 9-13, and 19-22.

Completion of tables, charts, and notes, as used in the test, requires transferring of information from the text. A validity table for the method is presented below as table 15.

Table 15: Analysis of justifications for method completion of tables as used within IELTS

Outcome: interpretation of the item as measures of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to read for specific relevant information, to read selectively, as well as read the whole text to organize the information under headings.	It requires familiarity with tables and charts, without which, reader may not be able to show reading ability, thus with construct-irrelevant difficulty. It may require writing, a contaminating factor for the inference of reading ability.	It may be considered an authentic task. It involves providing a few words, or phrases, not writing.

Is the item defensible? Yes. The item provides the reader with the more authentic task of extracting information and using the information in a similar way as required in an authentic situation of reading, e.g., to take notes, to summarize for future needs. It relies on the reading of the whole text and the development of its macrostructure in terms of the propositions and arguments, and requires specifically search reading, since the test taker will have to skim the text, and read carefully the parts of the text where the relevant information is. This item seems to have stronger arguments in favor of its

use to provide evidence converging for the inference of reading ability, in this case, the ability to read selectively and organize the information extracted from the text.

Method - Completion of summaries

Based on test 4, appendix 3, items 31-36.

Completion of summaries is somewhat different from completion of tables, charts, and notes and it is, thus, analyzed separately. Completion of summaries, as used in the test, requires filling the missing pertinent information in the gaps of a summary of the whole or part of the text. A validity table for the method is presented below as table 16.

Table 16: Analysis of justifications for the method completion of summaries within IELTS

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to read for global comprehension and to infer word meaning.	Comprehension of the source text and of the summary may have been attained without the reader's showing it by finding the right word for the gap in the text.	It provides evidence as to the ability to recognize vocabulary or infer meanings.

Is the item defensible? No. The task involves extracting, from the text, suitable words for gaps. It, thus, requires careful global reading of the text to be able find the words within their context and fitting them into the semantically and syntactically constraining gaps. However, getting some of the answers wrong may not mean that the reader was not able to read and understand the text. This item seems to have stronger

arguments against its use to provide evidence converging for the inference of reading ability.

Method - Choosing headings for paragraphs or sections of the text

Based on test 1, appendix 1, items 1-4, and 29-33.

Choosing headings for paragraphs of the text, as used in this test, requires finding the most suitable heading for each paragraph from a list of headings containing the headings and some distractors. A validity table for the method is presented below as table 17.

Table 17: Analysis of justifications for the method choosing headings within IELTS

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to establish the main idea of a paragraph.	Establishing paragraph main idea is different from selecting from a list. Words in the headings may be unknown, without context for inference of meaning.	Selecting from a list avoids the contaminating factor of writing. Words in the headings may be frequent, part of basic vocabulary required for good readers.

Is the item defensible? Yes. The item assesses comprehension of the main idea of paragraphs. Although reading is not choosing from a list, choosing from a list avoids the use of the ability of writing, which is a source of invalidity. In case the words in the headings are frequent and/or part of general vocabulary, this item will have stronger arguments in favor of its use to provide evidence converging for the inference of reading ability, in this case, the ability to recognize the main idea of paragraphs.

One item, number 3, requires considerations as to its technical quality, which is carried out in the technical quality table below as table 18.

Table 18. Analysis of technical quality of item 3, test 1, IELTS

Technical quality	Comments
Appropriateness of one answer	<p>According to the book, the key to item 3, for the heading of paragraph D, is letter (V), <i>the first rockets</i>. This is not the topic of the paragraph. The topic of the paragraph would be better reflected in <i>the first use(s) of rockets</i>, since it talks about the use of rockets as propellers of weapons.</p> <p>This is confirmed in the following paragraph taken from the text: <i>it was not until the eighteenth century that Europe became seriously interested in the possibilities of using the rocket itself as a weapon of war and not just to propel other weapons.</i></p>

Is the item defensible? Not the way developed. This might confuse the best readers, who may eventually choose the key by elimination, mostly like low proficient reader, or get the item wrong. The item, thus, has a low discrimination value. This item seems to have stronger arguments against its use to provide evidence converging for the inference of reading ability.

Method - Identification of a writer's claims, views or attitude

Based on test 1, appendix 1, items 18-21 and 36-39.

The item of *identification of a writer's claims, views, or attitudes* provides the reader with many statements with information not mentioned in the text, but related to it in terms of claims, views, and attitudes expected of the writer. A validity table for the method is presented below as table 19.

Table 19: Analysis of justifications for the method identifying writer’s claims, views or attitudes within IELTS

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to identify writer’s claims, views, and attitudes, part of the construct.	Being able to infer at this level may depend on the knowledge of the specific topic, adding construct irrelevance.	The item may be designed in a way to not depend on unshared or unmentioned knowledge.

Is the item defensible? Yes and no. The item assesses the ability to identify the writer’s claims, views, or attitude by asking the test taker to choose statements considered to reflect what is expected of the writer. It, thus, requires the identification of what is stated and also what is implied, involving search reading and inference making.

In case the item requires topical knowledge or specific information other than what is recoverable from the text, there may be no argument in favor of the use of the item. Looking into the instruction for the question *do the following statements reflect the claims of the writer in reading passage 2?*, comprised of questions 18 to 21, it can be argued that they are defensible on the grounds that they require inferencing based on the information recoverable from the text. That the item is limited to the information recoverable from the text is made clear in question number 21, which states *opponents of smoking financed the UCSF study*. The writer’s claim is *a more recent study by researchers at the University of California at San Francisco (UCSF) has shown that...* Since the answer is *Not given*, it is clear that there can be no inference as to who may have financed the study conducted by the UCSF researchers. This item may have stronger arguments in favor of its use to provide evidence converging for the inference

of reading ability, in this case, the ability to make inferences based on the information recoverable from the text.

Looking into the instruction for the question *do the following statements reflect the opinions of the writer in reading passage 3?*, comprised of questions 36 to 39, it can be argued that they are defensible on the grounds that they require inferencing based on, and limited to, the information recoverable from the text. Question 36 requires elementary inferencing of reinterpreting *says* in the question and *position taken* in the text, and leads to the conclusion that they have similar meaning. Question 37 requires elementary inferencing of reinterpreting the idea that the hypothesis is supported and retained until further test proves it incorrect, in the text, and the idea that the hypothesis is confirmed as true, in the question, and conclude that they are different.

Question 38 shows that the inferences to be made must be well supported by the information recoverable from the text. The statement in the question is *many people carry out research in a mistaken way*, whose answer, according to the book, is *not given*, meaning that there is no way to come to that conclusion based on the information given in the text.

However, there seems to be an argument that this statement reflects the writer's opinion, and that the answer to be given is *yes* rather than *not given*. The writer claims that there is a myth in scientific method that it is inductive, and his argument throughout the text is that scientific method is deductive rather than inductive. Based on his argument, it is plausible to infer that people might be carrying out research based on the inductive method, thus, in the writer's opinion, in a mistaken way. The answer, however, is that the information is not given.

This item may have stronger arguments in favor of its use to provide evidence converging for the inference of reading ability, in this case, the ability to make inferences based on the information recoverable from the text.

Method - Classification

Based on test 1, appendix 1, items 25-28.

In the item analyzed, the test taker must classify four statements based on the information given in the text, as findings, opinions, or assumptions of the studies mentioned in the text. A validity table for the method is presented below as table 20.

Table 20: Analysis of justifications for the method classification within IELTS

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to distinguish opinions, findings, and assumptions, to establish lexical cohesion, to draw elementary inferencing		

Is the item defensible? Yes. The item assesses an essential skill for any study reading, i.e., distinguishing opinions, findings, and assumptions. It focuses on *the study has shown, the report suggests, the report emphasizes*. It requires elementary inferencing, i.e., the recognition of propositions in the text and the matching with the propositions as the sentences in the item, involving identifying synonyms or synonymous sentences. It also requires the establishment of lexical cohesion, since the findings by the studies are reported using different lexical items (study, report). This item seems to have stronger

arguments in favor of its use to provide evidence converging for the inference of reading ability, in this case, the ability of developing propositions, using elementary inferencing to match with other propositions, and distinction of opinions, findings and assumptions.

Method - Matching phrases/lists

Based on test 1, appendix 1, items 11-14.

Matching phrases/lists, as used in this test, requires matching names representing specific things (projectiles) with drawing representing the things (projectiles). A validity table for the method is presented below as table 21.

Table 21: Analysis of justifications for the method matching phrases/lists within IELTS

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to establish a referent outside the text, i.e., non-textual referent.	It requires reinterpreting <i>0.7 meter</i> , in the text, to its representation in a scale, reinterpretation requiring specific knowledge of metric system, which is outside the construct of reading.	It may be considered an authentic task based on text reading. Metric system is part of the general knowledge of university students.

Is the item defensible? No. Although the task may be considered authentic, the test taker may understand the text without being able to show it through the item, since he/she may not know about the metric system. In this case, getting it wrong cannot be interpreted as lack of the ability of reading. This item seems to have stronger arguments against its use to provide evidence converging for the inference of reading ability.

General conclusions – IELTS examination

The examinations analyzed are more characterized in terms of integrative test than discrete-point²⁹, focusing more on comprehension than micro-skills. Many of the items will allow for collecting validity evidence converging for a valid interpretation of reading ability, based on the constructs used in this research, in that they provide evidence of the ability to read for global and local comprehension, based mostly on the information recoverable from the text, thus requiring skills such as vocabulary recognition, meaning inference of unknown words, parsing, cohesion establishment, and propositional inferences.

As to the threat of construct irrelevance, one item requires the knowledge of metric system, which is irrelevant for the construct of reading ability. As to the threat of construct underrepresentation, it is minimized with the diversity of methods that are used for each examination, diversity that allows the assessment of different skills of the construct.

There are also some technical problems (key and distractors) as shown, which may be sources of invalidity for the interpretation of reading ability, since they may not provide evidence for the test user to interpret ability based on the performance.

²⁹ Integrative testing is characterized in opposition to discrete point testing. While discrete point testing refers to testing one element at a time, integrative testing tests a combination of the language elements necessary for the completion of a task (Hughes, 2003).

5.2 – Analysis of the TOEFL test items

The TOEFL test is a proficiency test designed to measure the English language proficiency of nonnative speakers of English (TOEFL 2001 manual). It was first developed by Educational Testing Service (ETS) in 1963 (Pierce, 1992), and has been administered by the ETS until today.

The TOEFL test is characterized as having mostly multiple-choice questions, assessing different skills and levels of comprehension. The TOEFL 2000-2001 manual reads, “test items refer to what is stated or implied in the passage, as well as to words used in the passage” (p. 8). The items aim at assessing: 1) comprehension of main ideas; 2) inferences; 3) factual information stated in a passage; 4) pronoun referents; and 5) vocabulary (direct meaning, synonym, antonym) (p. 9). The analysis presented is based on the new version of the TOEFL test, and uses the construct accounting for the reading ability, as discussed in chapter 3, thus not necessarily considering congruence between what the specifications are and what the test contains. The analysis of the new version of the TOEFL test items follows exactly the order of items given in the test.

The title of the text used is “Opportunists and Competitors” and it is presented as Appendix 4. I present the parts of the text each item refers to, when possible. Thus, I present the item numbers, followed by the part of the text, then by the item proper. Item 12 requires the reading of the whole text. For this case, the text is given as Appendix 4.

Item 1

Item 1 is as follows:

Part of the text:

Growth, reproduction, and daily metabolism all require an organism to expend energy. The expenditure of energy is essentially a process of budgeting, just as finances are budgeted. If all of one's money is spent on clothes, there may be none left to buy food or go to the movies. Similarly, a plant or animal cannot squander all its energy on growing a big body if none would be left over for reproduction, for this is the surest way to extinction.

The word squander in the passage is closest in meaning to

- a) extend
- b) transform
- c) activate
- d) waste

A validity table for item 1 is presented below as table 22.

Table 22: Analysis of justifications for item 1 within TOEFL

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to recognize vocabulary or to infer its meaning. The target word <i>squander</i> is not a cognate word, and is not frequent ³⁰ , thus likely to be unknown, rendering the item text-dependent.	The key may be unknown, impeding test taker of showing recognition of target word, and inference making, and the key cannot be inferred, since it is given in isolation without context.	The distractors are cognate words. If test taker is able to infer the meaning of the target word, he/she will be able to choose the key by elimination. Also, the key is considered a frequent word ³¹ , and could be part of a threshold level for reading.

³⁰ This word is in the category of the least frequent in the Collins Cobuild dictionary.

³¹ This word is in the second top category of the most frequent in the Collins Cobuild dictionary

Is the item defensible? Yes. Since the context is rich in terms of information guiding inference making, the reader will most likely be able to infer the meaning of the target word. The information from *expenditure of energy* will give the reader some clues for the inference of the word *squander*. The rest of the information gives more clues for inference making. Getting this item right most likely means that the test taker was able to infer the meaning and was able to show that through the item. This item seems to have stronger arguments in favor of its use to provide evidence converging for the inference/to the interpretation of reading ability, in this case, the ability of meaning inferencing.

Item 2

Item 2 is as follows:

Part of the text:

Similarly, a plant or animal cannot squander all its energy on growing a big body if none would be left over for reproduction, for this is the surest way to extinction.

The word none in the passage refers to

- a) food
- b) plant or animal
- c) energy
- d) big body

A validity table for item 2 is presented below as table 23.

Table 23: Analysis of justifications for item 2 within TOEFL

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to establishing reference, an important skill in reading Establishing cohesion of <i>none</i> : - Ellipsis or elliptical expression: <i>none</i> and <i>energy</i>		

Is the item defensible? Yes. Establishing cohesion is in line with the constructs of language ability and reading ability used in this research. This item seems to have stronger arguments in favor of its use to provide evidence converging for the inference of reading ability, in this case, the ability of establishing cohesion.

Item 3

Item 3 is as follows:

Part of the text:

Growth, reproduction, and daily metabolism all require an organism to expend energy. The expenditure of energy is essentially a process of budgeting, just as finances are budgeted. If all of one's money is spent on clothes, there may be none left to buy food or go to the movies. Similarly, a plant or animal cannot squander all its energy on growing a big body if none would be left over for reproduction, for this is the surest way to extinction.

In paragraph 1, the author explains the concept of energy expenditure by

- identifying types of organisms that became extinct
- comparing the scientific concept to a familiar human experience
- arguing that most organisms conserve rather than expend energy
- describing the processes of growth, reproduction, and metabolism

A validity table for item 3 is presented below as table 24.

Table 24: Analysis of justifications for item 3 within TOEFL

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to read for the identification of how the writer develops and presents the arguments around one theme		

Is the item defensible? Yes. The test taker will read for local comprehension. Inference making is necessary for the idea of comparison, i.e., to compare the process of budgeting of the plants to what may be inferred as the process of budgeting as part of the human experience when deciding what to give priority when spending money, as a way of surviving the extinction. The *human* idea in the item is associated with *one's* in the text, since *one* may refer to an unknown human subject, and the possessive *'s* is mostly used referring to humans, association based on linguistic knowledge; the human experience in the items is associated with spending money in the text, association possible with some elementary inferencing. This item seems to have stronger arguments in favor of its use to provide evidence converging for the inference of reading ability, in this case, the ability of developing local comprehension, of identifying writer's intention of comparing, and of associating ideas based on inferencing.

Item 4

Item 4 is as follows:

Part of the text:

Almost all of an organism's energy can be diverted to reproduction, with very little allocated to building the body. Organisms at this extreme are 'opportunists'. At the other extreme are 'competitors', almost all of whose resources are invested in building a huge body, with a bare minimum allocated to reproduction.

According to the passage³², the classification of organisms as 'opportunists' or 'competitors' is determined by

- a) how the genetic information of an organism is stored and maintained
- b) the way in which the organism invests its energy resources
- c) whether the climate in which the organism lives is mild or extreme
- d) the variety of natural resources the organism consumes in its environment

A validity table for item 4 is presented below as table 25.

Table 25: Analysis of justifications for item 4 within TOEFL

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to read for local comprehension, establish cohesion, and draw inferences.		

Is the item defensible? Yes. The test taker will read for local comprehension. Establishment of cohesion in terms of the identification of ellipsis is the first step to recognize the noun of *little* after *all of energy*. The identification of three words – *investing*, *diverting*, and *allocating* as synonymous in the context is the second step. Inference making is necessary for reinterpreting *all of energy diverted to reproduction*,

³² Although they use passage rather than a specific paragraph, the information for the answer can be found in the paragraph given previously, paragraph 3 in the text.

with very little allocated to building the body in the text as the way of investing energy resources in the item. This item seems to have stronger arguments in favor of its use to provide evidence converging for the inference of reading ability, in this case, the ability of establishment of cohesion, recognition of synonyms, and association of superordinate proposition in the item to their subordinate propositions in the text.

Item 5

Item 5 is as follows:

Part of the text:

Dandelions are good examples of opportunists. Their seedheads raised just high enough above the ground to catch the wind, the plants are no bigger than they need be, their stems are hollow, and all the rigidity comes from their water content. Thus, a minimum investment has been made in the body that becomes a platform for seed dispersal. These very short-lived plants reproduce prolifically; that is to say they provide a constant rain of seed in the neighborhood of parent plants.

The word dispersal in the passage is closest in meaning to

- a) development
- b) growth
- c) distribution
- d) protection

A validity table for item 5 is presented below as table 26.

Table 26: Analysis of justifications for item 5 within TOEFL

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to recognize vocabulary or infer its meaning The target word <i>dispersal</i> is not frequent ³³ , thus likely to be unknown, good choice for inference making.	The target word is rather cognate. Matching it with the key may be done without the text, rendering the item as passage-independent.	Test taker may need confirmation from the text.

Is the item defensible? No. Although the target word is infrequent, the context is good in terms of information guiding inference making, and the key is a cognate word not impeding the demonstration of the inference of the target word, the target word itself being a cognate. Getting this item right most likely means matching the target word with the key based on the knowledge of Portuguese, with no need for reference to the text. This only shows vocabulary recognition. This item seems to have stronger arguments against its use to provide evidence converging for the inference of reading ability, since it may be a passage-independent item.

Item 6

Item 6 is as follows:

Part of the text: highlighted sentence

These plants are termed opportunists because they rely on their seeds' falling into settings where competing plants have been removed by natural processes, such as along an eroding riverbank, on landslips, or where a tree falls and creates a gap in the forest canopy.

³³ This word is in the category of the least frequent in the Collins Cobuild dictionary.

Which of the sentences below best expresses the essential information in the highlighted³⁴ sentence in the passage? *Incorrect* choices change the meaning in important ways or leave out essential information

- a) Because their seeds grow in places where competing plants are no longer present, dandelions are classified as opportunists
- b) Dandelions are called opportunists because they contribute to the natural processes of erosion and the creation of gaps in the forest canopy.
- c) The term opportunists applies to plants whose seeds fall in places where they can compete with the seeds of other plants
- d) The term opportunists applies to plants whose falling seeds are removed by natural processes

A validity table for item 6 is presented below as table 27.

Table 27: Analysis of justifications for item 6 within TOEFL

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to read for local comprehension, establish cohesion and draw inference.		

Is the item defensible? Yes. Establishment of cohesion in terms of the lexical cohesion *dandelions* and *these plants* is the first step. The task requires elementary inferencing: the recognition of *dandelions are classified as opportunists* in the item, as a reorganization or reinterpretation of *these plants are termed opportunists* in the text; and the recognition of *their seeds grow in places where competing plants are no longer present* in the item, as a reorganization or reinterpretation of *they rely on their seeds' falling into settings where competing plants have been removed* in the text. This item seems to have stronger arguments in favor of its use to provide evidence converging for

³⁴ In the appendix, rather than highlighted, the sentence is underlined.

the inference/to the interpretation of reading ability, in this case, the ability of establishing cohesion and drawing elementary inferencing.

Item 7

Item 7 is as follows:

Part of the text:

An oak tree is a good example of a competitor. A massive oak claims its ground for 200 years or more, outcompeting all other would-be canopy trees by casting a dense shade and drawing up any free water in the soil.

The word massive in the passage is closest in meaning to

- a) huge
- b) ancient
- c) common
- d) successful

A validity table for item 7 is presented below as table 28.

Table 28: Analysis of justifications for item 7 within TOEFL

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to infer word meaning of the target word <i>massive</i> , since it is not exactly a cognate word, and not exactly frequent ³⁵ , thus likely to be unknown.	The grammatical structure of the sentence <i>outcompeting all the would-be canopy trees</i> is complex. The word <i>casting</i> is infrequent, and <i>shade</i> is field-specific. The key is rather infrequent in formal situations such as academic texts.	Being able to infer the target word meaning using the context shows high level of proficiency, and provides high discrimination values.

³⁵ This word is in the category 3 of the scale ranging from 1 to 5 in the Collins Cobuild Dictionary.

Is the item defensible? No. The context is fairly rich in terms of information guiding the inference making. The information from the context for inference comes from *outcompeting all the would-be canopy trees* and *casting dense shade*, which seems rich enough for inferencing. Getting this item right most likely means that the test taker was able to infer the meaning, making use of the context, thus showing knowledge of more complex grammar, and the recognition of the phrase *casting shade*. It will also mean that test taker knew the word *huge*, or used the strategy of elimination, based on the knowledge of the distractors, or on the fact that the distractors are cognate words. However, the reader may not be able to show that ability of inference making through the item, because of the lack of knowledge of the word *huge*, which is more frequently used in informal situations, thus not reflecting the needs of the criterion. There seems to be stronger argument against the use of this item to provide evidence converging for the inference/to the interpretation of reading ability, since the item may become a vocabulary item.

Item 8

Item 8 is as follows:

Part of the text:

The opposite of an opportunist is a competitor. These organisms tend to have big bodies, are long-lived, and spend relatively little effort each year on reproduction. An oak tree is a good example of a competitor. A massive oak claims its ground for 200 years or more, outcompeting all other would-be canopy trees by casting a dense shade and drawing up any free water in the soil. The leaves of an oak tree taste foul because they are rich in tannins, a chemical that renders them distasteful or indigestible to many organisms. The tannins are part of the defense mechanism that is essential to longevity. Although oaks produce thousands of acorns, the investment in a crop of acorns is small compared with the energy spent on building leaves, trunk, and roots. Once an oak tree becomes established, it is likely to survive minor cycles of drought and even fire. A population of oaks is likely to be relatively stable through time, and its survival is likely to depend

more on its ability to withstand the pressures of competition or predation than on its ability to take advantage of chance events. It should be noted, however, that the pure opportunist or pure competitor is rare in nature, as most species fall between the extremes of a continuum, exhibiting a blend of some opportunistic and some competitive characteristics.

All of the following are mentioned in paragraph 7 as contributing to the longevity of an oak tree EXCEPT

- a) the capacity to create shade
- b) leaves containing tannin
- c) the ability to withstand mild droughts and fire
- d) the large number of acorns the tree produces

A validity table for item 8 is presented below as table 29.

Table 29: Analysis of justifications for item 8 within TOEFL

Outcome: interpretation of the item as measures of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to read for more global comprehension. The task requires integrating information across the text, which demands inferencing around the concept of longevity.	The item is best answered with the knowledge of specific topic, adding to construct irrelevance. Inferencing is compromised due to poor knowledge of plants and trees in general, and oak tree in particular. Negative statements may confuse test takers who may choose the first correct answer, not the key, which contains wrong information.	Reading several times may provide more information for inferencing making, even for the test takers unfamiliar with the topic. The instruction that the key is the only information not mentioned in the text is signaled with the capitalized word <i>except</i>

Is the item defensible? No. The item requires the ability to integrate information across the text. The unfamiliar topic may create construct-irrelevant difficulty for the integration process. It may have, thus, low discrimination index, since it may be confusing for the test takers with high proficiency in language, causing them to choose

the wrong option, and so difficult for the test takers with low proficiency that they would use their 25% correct chance for guessing the key. This item seems to have stronger arguments against its use to provide evidence converging to the inference of reading ability.

Item 9

Item 9 is as follows:

Part of the text:

The opposite of an opportunist is a competitor. These organisms tend to have big bodies, are long-lived, and spend relatively little effort each year on reproduction. An oak tree is a good example of a competitor. A massive oak claims its ground for 200 years or more, outcompeting all other would-be canopy trees by casting a dense shade and drawing up any free water in the soil.

According to the passage, oak trees are considered competitors because

- a) they grow in areas free of opportunists
- b) they spend more energy on their leaves, trunks and roots than on their acorns
- c) their population tends to increase or decrease in irregular cycles
- d) unlike other organisms, they do not need much water or sunlight

A validity table for item 9 is presented below as table 30.

Table 30: Analysis of technical quality for item 9 within TOEFL

Technical quality	Comments
Appropriateness of the options	<p>There may be two possible answers here.</p> <p>Characteristics of competitors: they have big bodies, are long-lived, and spend relatively little effort each year on reproduction.</p> <p>There is a trade off of energy expenditure.</p> <p>If on reproduction, not on body</p> <p>If on body, not on reproduction</p> <p>Competitor on body, therefore not on reproduction (acorns).</p> <p>Option (b) <i>they spend more energy on their leaves, trunks and roots than on their acorns</i> is correct.</p> <p>If big body, dense shade</p> <p>If dense shade, the area becomes free of opportunists/no opportunists</p> <p>Option (a) <i>they grow in areas free of opportunists</i> may be inferable.</p>

Is the item defensible? No. Although the key, letter b, is a more suitable answer for the item, especially considering that it has more of the characteristic of a definition as required by the word *considered* in the stem, the item may be confusing for the more proficient test taker who might get it wrong in spite of being able to understand the text. The item, then, becomes non-dependable for the inference of language ability, since it may provide low discrimination index between those test takers with the ability and those without the ability. This item seems to have stronger arguments against its use to provide evidence converging for the inference of reading ability.

Item 10

Item 10 is as follows:

Part of the text
Same as item 08

In paragraph 7, the author suggests that most species of organisms

- a) are primarily opportunists
- b) are primarily competitors
- c) begin as opportunists and evolve into competitors
- d) have some characteristics of opportunists and some of competitors

A validity table for item 10 is presented below as table 31.

Table 31: Analysis of justifications for item 10 within TOEFL

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to comprehend a text locally. The task requires elementary inferencing, since reorganization or reinterpretation of the information in the text is necessary.		

Is the item defensible? Yes. The reader will have to associate some *characteristics of opportunists and some of competitors* in the item with the *pure opportunist or pure competitor is rare in nature, as most species fall between the extremes of a continuum, exhibiting a blend of some opportunistic and some competitive characteristics* in the text, requiring elementary inferencing. This item seems to have stronger arguments in favor of its use to provide evidence converging for the inference of reading ability, in this case, the ability to read for local comprehension and draw elementary inferencing.

Item 11

Item 11 is as follows:

Part of the text:

Opportunists must constantly invade new areas to compensate for being displaced by more competitive species. Human landscapes of lawns, fields, or flowerbeds provide settings with bare soil and a lack of competitors that are perfect habitats for colonization by opportunists. ■ Hence, many of the strongly opportunistic plants are the common weeds of fields and gardens. ■ Because each individual is short-lived, the population of an opportunist species is likely to be adversely affected by drought, bad winters, or floods. ■ If their population is tracked through time, it will be seen to be particularly unstable—soaring and plummeting in irregular cycles. ■

Look at the four squares [■] that indicate where the following sentence could be added to the passage.

Such episodic events will cause a population of dandelions, for example, to vary widely.

Where would the sentence best fit?

A validity table for item 11 is presented below as table 32.

Table 32: Analysis of justifications for item 11 within TOEFL

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to establish lexical cohesion. Establishing lexical cohesion: <i>such episodic events</i>	Inserting does not equal reading	Inserting may avoid unintended clues given as the options

Is the item defensible? Yes. In spite of the fact that inserting a sentence does not equal reading, the evidence elicited through this item will lead to the interpretation that the test taker is able to establish cohesion. In this case, the lexical cohesion is

established if the reader knows the meaning and function of *such* as a reference word, and knows that event is one of those words to be lexicalized within the context, being what Nuttall (1996) has called text-structuring words. *Such events*, in the case, refers back to *drought*, *bad winters*, or *floods*. This item seems to have stronger arguments in favor of its use to provide evidence converging for the inference of reading ability, in this case, the ability to establish lexical cohesion.

Item 12

Item 12 is as follows:

Part of the text:

Whole text (see appendix 4)

Directions: Complete the table by matching the phrases below

Directions: Select the appropriate phrases from the answer choices and match them to the type of organism to which they relate. TWO of the answer choices will NOT be used. ***This question is worth 4 points.***

Drag your answer choices to the spaces where they belong. To remove an answer choice, click on it. To review the passage, click on **View Text**.

Answer Choices		Opportunists
Vary frequently the amount of energy they spend in body maintenance	●	
Have mechanisms for protecting themselves from predation	●	
Succeed in locations where other organisms have been removed	●	
Have relatively short life spans	●	
Invest energy in the growth of large, strong structures	●	
Have populations that are unstable in response to climate conditions	●	Competitors
Can rarely find suitable soil for reproduction	●	
Produce individuals that can withstand changes in the environmental conditions	●	
Reproduce in large numbers	●	

A validity table for item 12 is presented below as table 33

Table 33: Analysis of justifications for item 12 within TOEFL

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to read the whole text and recognize some of its main propositions.	It may be assessing recall rather than comprehension of detail information	It is the last item and test taker will probably remember most of the information, and may reread otherwise.
	Elementary inferences are required in terms of use of different, though synonymous words in the <i>answer-choices</i> column in the item.	It requires the knowledge of words other than the ones in the text, such as <i>span</i> .	They are sentences, giving test takers the chance for inference making of meanings.

Is the item defensible? Yes. This is the only item in the test which is not multiple-choice. It can be considered a matching activity, or a set of true or false activity, where each one will be true for one type of organism and necessarily false for the other.

It assesses the ability to develop a macrostructure of the text in terms of the main ideas, and in terms of recognizing the rhetorical organization of a comparison/contrast structure. A potential problem with the item is that, since it requires recall of detail information, it may be assessing what test taker can remember of the text. Comprehension as the outcome of reading does not equate remembering (Urquhart & Weir, 1998). Rereading the text allows test taker to provide the answer without the need to store detail information. This item seems to have stronger arguments in favor of its use to provide evidence converging for the inference of reading ability, in this case, the ability to draw elementary inferencing, and to develop arguments around the two themes, and the macrostructure of the text.

Analysis of items of TOEFL practice tests

The items analyzed in this section are taken from practice books for the TOEFL tests. They have been included for the analysis due to their relevance for the discussion, in particular, on items providing construct-irrelevant evidence. Since they are selected from various sources, they are numbered as a sequence of the previous items, thus not numbered based on their original numbers.

Item 13 appendix 5

In ETS (1995, p. 30).

This item is as follows:

Part of the text:

...she also photographs away from her studio at various architectural sites, bringing camera, lights, mirrors, and a crew of assistants to transform the site into her own abstract image. (cont.)

The word transform in line 6 is closest in meaning to

- a) move
- b) extend
- c) change
- d) interpret

A validity table for item 13 is presented below as table 34.

Table 34: Analysis of justifications for item 13 within TOEFL

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to infer the meaning of a word by the use of the context	The target word <i>transform</i> does not require meaning inference from Brazilians since it is a cognate word. The item becomes a vocabulary item, since Brazilian test takers will have to know the words as options without further context.	The word <i>change</i> – the key – may be considered part of the basic vocabulary, since it is very frequent. The dictionary Collins Cobuild characterizes it as very frequent, category that accounts for approximately 75% of all English usage.

Is the item defensible? No. This is a typical example of an item which does not show reading ability, that is, the reader may have understood its context without being able to show it through the item. The target word is a cognate. There can be no argument that its meaning was not accessed by Brazilian test takers. The key, however, is not a cognate, and is given in isolation as options, without context. This item becomes, thus, a vocabulary item, aimed at assessing vocabulary knowledge. Since the key is a very frequent word, there can be the argument that it is part of a threshold of basic vocabulary, necessary for reading and for the use of strategies such as meaning inference. The outcome of the item must come, however, as a construct of basic vocabulary knowledge, which is different from the construct of reading ability. This item seems to have stronger arguments against its use to provide evidence converging for the inference of reading ability.

Item 14 appendix 6

In ETS (1995, p. 32).

This item is as follows:

Part of the text:

...In the core of the Sun, the pressures are so great against the gases that... (cont.)

The word great in line 4 is closest in meaning to

- a) dangerous
- b) unknown
- c) variable
- d) strong

A validity table for item 14 is presented below as table 35.

Table 35: Analysis of justifications for item 14 within TOEFL

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to infer word meaning.	The key <i>strong</i> may not be known to the test taker. Comprehension of the text may have occurred without the possibility of showing it. It becomes a vocabulary item.	The key <i>strong</i> is considered in the dictionary Collins Cobuild very frequent, category that accounts for approximately 75% of all English usage.

Is the item defensible? No. The target word *great* is not a cognate word, thus a good choice for the item. Although it is a frequent word according to Collins Cobuild dictionary, its meaning in this context is more infrequent. This may require inference making within its context, possible with the presence of the word *pressure* which collocates with the key *strong*. If the key is not known, this item becomes, for Brazilian

test takers, a vocabulary item, aimed at assessing vocabulary knowledge of the key, given in isolation, without context. Since the key is a very frequent word, the same argument as item 14 above can be used. As to its outcome, it must also come as the construct of basic vocabulary knowledge, not of reading ability. The way it is developed, this item seems to have stronger arguments against its use to provide evidence converging for the inference of reading ability.

Item 15 appendix 7

In Rymniak and Shank (2002, p. 425)

This item is as follows:

Part of the text:

One of the most renowned Spanish architects of all time was Antoni Gaudi. Gaudi's emergence as one of the Spain's prominent artists at the end of the nineteenth century marked a milestone in the art world. (cont.)

Antoni Gaudi's fame is due primarily to his world-famous

- a) paintings
- b) architectural structures
- c) political skills
- d) business acumen

A validity table for item 15 is presented below as table 36.

Table 36: Analysis of justifications for item 15 within TOEFL

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to find specific information, an important ability for reading in a foreign language.	It may be answered based on background knowledge, without reading the text. It may also be answered by matching words only, <i>architectural structures</i> in the item and <i>architects</i> in the text.	It requires reading of the item.

Is the item defensible? No. Since this item may be answered based on background knowledge or the strategy of matching words rather than reading the text, it seems to have stronger arguments against its use to provide evidence converging for the inference of reading ability.

General conclusions – TOEFL test

The TOEFL items analyzed show that it is mostly a monomethod, multiskill, multilevel test. It is generally construct representative. Item 1 provides evidence about the inferential process based on local comprehension, and item 2, about referential process, both requiring integrating information. Also, item 6 provides evidence of local comprehension and of elementary inferencing to associate propositions in the text with propositions as options.

There is however, construct irrelevance, such as item 5, which can be answered without reference to the text, mostly based on the knowledge of Portuguese, not the

target language. It is also the case of some items focusing on vocabulary knowledge rather than the ability to read a text.

This is a problem for the collection of evidence for inference of reading ability, since assessing vocabulary knowledge is assessing a construct other than reading ability, particularly because language ability and reading ability have a strategic component, as described in chapter 3, which makes it possible for the meaning of unknown words to be inferred based on the context.³⁶

5.3 – Analysis of the items used by EAP Teacher 1

The analysis involved three tests, which do not follow the same pattern in terms of the number of items and methods used. The test analyzed (see appendix 8) to be presented here is not representative of all tests used by the teacher. It is made up of six items, with different methods and various skills, so it is a multimethod multiskill test. Some of the instructions are not clear as to how to respond to the item. Although this is not usually a problem for classroom tests because teachers may assist the test takers in case of doubt, it has caused some uncertainty for the analysis in this research.

Item 1

Item 1 is as follows:

Look at the sentences below. All the words in italics are nonsense words. Discover what these words mean from the context of the sentence. Sometimes more than one word is possible.

1 – It was a very cold day, so I put a *tribbet* around my neck.

³⁶ I recognize, however, that the assessment of vocabulary is not incongruent with their specifications, since it is claimed in their manual that the questions in the reading section will also assess vocabulary (direct meaning, synonym, antonym) (TOEFL 2001 manual, p. 9).

- 2 – He was so *fliglive* that he drank a whole bottle of Coke.
 3 – Mary did three tralets yesterday but failed them all because she hadn't studied enough
 4 – She did the exam very *trodly* because she had a headache.
 5 – The doctor *sarked* very late at work because he overslept.

A validity table for item 1 is presented below as table 37.

Table 37: Analysis of justifications for item 1 EAP teacher 1

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence of the ability to understand and use context to infer meaning of unknown words.	The words are in loose sentences, and inferences will be at the local level. Words are inexistent.	Inferences at the local level are important. This allows the assessment of the ability to make inference rather than word knowledge.

Is the item defensible? Yes. The fact that they are nonsense words will allow inferences to be made concerning to ability assessed, with no possibility of being recognized as a familiar word or a cognate word. Thus, this item seems to have stronger arguments in favor of its use to provide evidence converging to the inference of reading ability, in this case, the ability of making use of the context to infer meaning of unknown words.

Item 2

Item 2 is as follows:

Complete the gaps in the text with the correct words

Americans are well-known for being If we're taking a in the park and we pass someone, we usually say *hi!* or *how's it going?* to each And we usually say a few words to people in stores, bars, and banks. But remember:is not friendship: it's In the United States, it's just as to make real friends as it is anywhere else.

hard – politeness – friendly – walk – other – friendliness

A validity table for item 2 is presented below as table 38.

Table 38: Analysis of justifications for item 2 EAP teacher 1

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence as to the ability to use semantic and syntactic constraints and to choose the word semantically and syntactically appropriate for the gap.	Gaps are not very semantically constraining, and the first sentence already has a gap with little context. Providing the words means the direction from meaning to the words, which is the direction of the productive skills. Reading is not a productive skill. Providing the words implies necessarily knowing the meaning of the words available in a list. It becomes a vocabulary item.	Gaps are constraining enough for the choice of words in the bank of words. Providing the meaning is the same direction as inference making.

Is the item defensible? No. Although the item assesses the ability to make use of the constraints of the context for inference making, the possibility of the test taker not knowing the words in the list renders this item a vocabulary item. There seems to be

stronger arguments against the use of this item to provide evidence converging to the inference of reading ability.

Item 3

Item 3 is as follows:

What are the articles related to these headlines?

- 1 – Teenagers say AIDS is their biggest fear
- 2 – World champion swimmer suspended after drug test
- 3 – Explosion kills 20 people

A validity table for item 3 is presented below as table 39.

Table 39: Analysis of justifications for item 3 EAP teacher 1

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence that test taker is able to use linguistic information to infer the topic related to the headlines which helps activating the right schemata for comprehension.	Text is too short, just one sentence. Vocabulary may be unknown and with little context for inference. Information processed in the sentence may not lead to the topic because test taker does not have the schemata.	Text is long enough for the purpose of identifying the topic. Vocabulary may have been worked with in the classroom. Topic may have been worked with in the classroom.

Is the item defensible? No. Finding the topic relies on background knowledge, and background knowledge is not part of the construct used in this research. Evidence collected through the item may not be used for inference about language or reading

ability. This item seems to have stronger arguments against its use to provide evidence converging to the inference of reading ability.

Item 4

Item 4 is as follows:

Match the second part of each sentence

- 1 – I speak fluent German, () but I enjoy dancing.
 2 – we aren't going to Germany, () and knows many good restaurants.
 3 – I don't do any sports, () so there's no need to buy tickets.
 4 – Steven eats out a lot with friends () but there were some strange people in the restaurant.
 5 – At first, everything seemed fine, () and I've just come back from Germany.

A validity table for item 4 is presented below as table 40.

Table 40: Analysis of justifications for item 4 EAP teacher 1

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence test taker is able to establish cohesion.	Reading is not matching sentences.	Reading is establishing relations.

Is the item defensible? Yes. This item seems to have stronger arguments in favor of its use to provide evidence converging to the inference of reading ability, in this case establishment of cohesion.

Item 5

Item 5 is as follows:

Read the text below and give three reasons why flying is bad for people's health

A validity table for item 5 is presented below as table 41.

Table 41: Analysis of justifications for item 5 EAP teacher 1

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence test taker is able to identify the information around the theme of the text.	It involves writing ³⁷ .	It involves identifying the information in the text and copying the words, not writing.

Is the item defensible? Yes. Writing to some degree has sometimes been involved in the assessment of reading, which has been considered a contaminating factor. This item involves writing to the extent of copying words only, having, therefore, the advantage that options are not given as it is done in the case of MCQ, hence minimizing the possibility of guessing. It, then, seems to have stronger arguments in favor of its use to provide evidence converging to the inference of reading ability, in this case, search reading to find specific information.

Item 6

Item 6 is as follows:

Explain the following compound nouns from the text: heart attack, economy class, leg room, time zone, blood pressure.

³⁷ The test is not clear as to whether it is to be answered in the native or in the target language.

A validity table for item 6 is presented below as table 42.

Table 42: Analysis of justifications for item 6 EAP teacher 1

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence as to the ability to parse correctly in English, that is, to identify the head and the modifier in the phrase, and, in case the words are unknown, be able to infer meanings through the context.	Some words within the phrases are cognates, not requiring context. The cognate word <i>zone</i> in <i>time zone</i> does not help for this situation, and the context in which it is inserted does not help either. Test taker may be able to parse, but not explain because context is not enough. So <i>time zone</i> is vocabulary sub-item.	This is not necessarily a word inferencing item, but a parsing item. Cognates are not a problem.

Is the item defensible? Yes. This type of item is probably designed within the perspective of contrastive analysis. Any teacher working with Brazilians learning English will know that the recognition and understanding of the constituents of a noun phrase are problems for them, since it is a situation in which both languages differ. Assessing this ability may, thus, be part of tests developed to assess the reading ability of Brazilians. Item 6 seems to have stronger arguments in favor of its use to provide evidence converging to the inference of reading ability, in this case, establishing syntactical relations.

General conclusions – EAP teacher 1

The items analyzed show that it is mostly a multimethod, multiskill test. It is generally construct representative. Most items focus on micro-skills, such as inference of word meaning, establishment of cohesion and of syntactical relation for the parsing, and there is one item requiring reading the whole text to find specific information.

There is, however, construct irrelevance, such as the item (item 3) focusing on the headlines, and the item which requires knowledge of words given in the list (item 2). Also, there is the problem, for reliability, of having only 6 items, which may not be considered enough to collect evidence for the interpretation of reading ability.

5.4 – Analysis of the items used by EAP teacher 2

The test analyzed (appendix 9) comprises 15 items, 10 of which are multiple-choice questions, 04 are short-answer questions, and 01 is a translation task. It may be considered to have different methods and various skills, being a multimethod multiskill test. Since the test is very similar to the TOEFL tests, I will only choose some of the items deserving analysis. All the questions, stems and options are in Portuguese, and will be translated into English by this researcher.

Item 2

Item 2 is as follows:

Part of the text:

Luisa May Alcott, an American author best known for her children's books *Little Women*, *Little Men*, and *Jo's Boys*, was profoundly influenced by her family, particularly her father. (cont.)

Na linha 2, a palavra “particularly” assemelha-se mais em termos de significado a (the word ‘particularly’ in line 2 is closest in meaning to):

- a) parcialmente por (partially for)
- b) estranhamente (strangely)
- c) exceto por (except for)
- d) especialmente (particularly)

A validity table for item 2 is presented below as table 43.

Table 43: Analysis of justifications for item 2 EAP teacher 2

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to infer the appropriate meaning or make use of the context for inference meaning.	The word is a cognate word, and the options are given in Portuguese. Matching based on the knowledge of Portuguese is enough. It becomes a passage-independent item.	The test taker may not know it is a cognate word and may need to resort to the text for confirmation.

Is the item defensible? No. Even if there is the need for confirmation, the very fact that the test taker may use guessing strategy to get the item right will not allow inferences about reading ability, and the test user will not know whether this shows ability or just guessing. What is interesting about this item is that it was used by a Brazilian teacher with Brazilian students. The choice of the word *particularly* suggests that the teacher was not certain what to focus on for the assessment of reading ability, or how to develop an item for inference making. This item seems to have stronger arguments against its use to provide evidence converging to the inference of reading ability.

Item 4

Item 4 is as follows:

Part of the text:

...She was the daughter of Bronson Alcott, a well-known teacher, intellectual, and free thinker who advocated abolitionism, women's rights, and vegetarianism long before they were popular. He was called a man of unparalleled intellect by his friend Ralph Waldo Emerson. Bronson Alcott instilled in his daughter his lofty and spiritual values and in return was idolized by his daughter. (cont.)

Na linha 5, a palavra “lofty” assemelha-se mais em termos de significado a (the word “lofty” in line 5 is closest in meaning to)

- a) comum (common, ordinary)
- b) generoso (generous)
- c) egoísta (selfish)
- d) simpático (nice)

A validity table for item 4 is presented below as table 44.

Table 44: Analysis of justifications for item 4 EAP teacher 2

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to infer the meaning of the target word by making use of the context and by relating information.		

Is the item defensible? Yes. Although *lofty* does not mean exactly generous, generous might be considered the closest meaning. This item makes use of the infrequent word, *lofty*, unlikely to be known by the test takers, thus requiring the use of the text, the use of the information of the context for the inference of its meaning, and the skill of relating scattered information for inference making. In addition to that,

context is rich in providing cues for inference of the meaning of the target word. This item seems to have stronger arguments in favor of its use to provide evidence converging to the inference of reading ability, in the case, the use of context for the inference of unknown words.

Item 6

Item 6 is as follows:

Part of the text:

As a result, Luisa had to begin helping to support her family at a young age, by taking a variety of low-paying jobs as a seamstress, a maid, and a tutor.

Qual das seguintes atividades NÃO foi exercida por Luisa para ganhar dinheiro em sua juventude (which of the following jobs did Luisa NOT take to earn her living at a very young age)?

- a) trabalhou como costureira (worked as a seamstress)
- b) trabalhou como faxineira (worked as a maid)
- c) trabalhou como professora (worked as a tutor, teacher)
- d) trabalhou em uma loja (worked at a store)

A validity table for item 6 is presented below as table 45

Table 45: Analysis of justifications for item 6 EAP teacher 2

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to find specific information.	Test taker may be able to find the information but may not be able to answer, because does not know the meaning of the words <i>seamstress</i> and <i>maid</i> .	The context will provide the information that seamstress and maid are low-paying jobs, providing the test taker with some information as to the jobs.

Is the item defensible? No. Although finding and understanding specific information is probably the aim of this item, it has become a vocabulary item. For test takers to answer this question, they will have to know the meanings of the job-related words in the text. Although the context provides information that they are low-paying jobs, the test taker will know it is not letter c, because *tutor* in the text has a cognate word in Portuguese *tutora*, with *professora* as a synonym in Portuguese. The test taker will know this is a job taken by Luisa. However, the test taker will not be able to choose the key among the other three options which are all low-paying jobs³⁸. The only way of getting this item right is by knowing the meaning of the words *maid* and *seamstress* in advance, since the context does not provide enough clues for the inference of their meanings. This becomes, thus, a vocabulary item. This item seems to have more arguments against its use to provide evidence converging to the inference of reading ability.

Item 8

Item 8 is as follows:

Part of the text:

With the success of this novel she was able to provide for her family, giving her father financial security that until then he had never experienced.

Pode-se inferir a partir do texto que Luisa May Alcott usou o sucesso de *Little Women* para (It is possible to infer from the text that Luisa May Alcott used the success of *Little Women* to)

- a) presentear-se com tudo o que sempre quis (to buy herself all the presents she always wanted)
- b) atingir sucesso financeiro e pessoal (attain personal and financial success)
- c) dar a seu pai uma prova intangível de seu amor (give her father intangible proof of her love)

³⁸ In case *tutor* was not a cognate word, pragmatic inference would tell the test taker to choose letter c, *worked as a teacher*, the only real low-paying job today.

- d) separar-se de sua familia (separate from her family)

A validity table for item 8 is presented below as table 46.

Table 46: Analysis of technical quality for item 8 EAP teacher 2

Technical quality	Comments
Appropriateness of the key	<p>Is it possible to infer from the text that Louisa gave her father an intangible proof of her love by giving him financial security?</p> <p>Since this father has lofty and spiritual values, inference making allows the claim that intangible proof of love for such a man who advocates women's rights and vegetarianism is to engage in fights for women's rights, or fights against the killing of animals to feed human beings.</p>

Is the item defensible? No. Although option 'c' is the most likely choice as the key by the most proficient reader, the strategy of elimination may be the way to get the item right. Since there is the argument that inference making allows for more than one choice, it becomes a trick item, in which a proficient test taker may disagree with the key and try another option. The item may have a low discrimination value. This item seems to have stronger argument against its use to provide evidence converging to the inference of reading ability.

Item 9

Item 9 is as follows:

Part of the text

Whole text (see appendix 9)

O propósito do autor nessa passagem é (the purpose of the author in the passage is):

- explicar como a autora tornou-se famosa (explain how the writer became famous)
- descrever a influencia da família na vida da escritora (describe the influence of the family on the life of the writer)

- c) apoiar as teorias educacionais de Bronson Alcott (support the educational theories by Bronson Alcott)
- d) mostrar o sucesso que pode ser atingido por um(a) autor(a) (show the success a writer can achieve)

A validity table for item 9 is presented below as table 47.

Table 47: Analysis of justifications for item 9 EAP teacher 2

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to develop the macrostructure of the text, with the main idea and the writer's purpose with the article (functional knowledge).	Establishing functional knowledge requires pragmatic knowledge.	It requires pragmatic knowledge at the level of functional knowledge, which is implied in the text. Also, the functions of explaining, describing, showing are part of the ideational function, characteristics of expository text.

Is the item defensible? Yes. This item shows evidence of global comprehension and of inference making for the understanding of the author's purpose, hence, assessing the establishment of the functions of the language (explain, describe, give support, show), which are necessary for the coherence of the text. Although it involves pragmatic knowledge, it can be argued that the level of inference making is constrained by what is implied by the writer. The plausible criterion, as suggested by Pearson and Johnson (1978), to render the answer as having a bearing on the text is based on pragmatic knowledge in terms of functional knowledge, as described by Bachman (1990), and Bachman and Palmer (1996). Using the typology by Pearson and Johnson (1978) describe in chapter 3, this question may be considered a textually-implicit question. This item seems to have stronger arguments in favor of its use to provide

evidence converging to the inference of reading ability, in the case, the use of the text to develop the macrostructure of the text.

Item 11

Item 11 is as follows:

Retire do texto palavras formadas por prefixação e duas formadas por sufixação e suas respectivas paráfrases³⁹ (Identify in the text two words with suffixes and two words with prefixes and their respective paraphrases)

A validity table for item 11 is presented below as table 48.

Table 48: Analysis of justifications for item 11 EAP teacher 2

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that test taker knows rules of word formation and is able to decode words and access their meanings.	This requires knowledge of metalanguage – suffixes and prefixes.	Metalanguage has probably been used during the EAP classes.

Is the item defensible? Yes. Knowledge of word formation is essential to help the reader to access meanings of familiar words and infer meanings of unknown words. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, knowledge of prefixes and suffixes, thereby awareness of word formation, which contributes to the lower level processing of word decoding and lexical access.

³⁹ The instruction is not clear as what to do. Based on what I can understand, there seems to be two problems. One refers to the definition of paraphrase, which seems to mean here the corresponding phrase in Portuguese, conclusion also based on item 13. The second refers to the fact that the corresponding phrases are not to be identified in the text, but are provided by the test taker.

Item 12

Item 12 is as follows:

Retire do texto uma conjunção de resultado e uma conjunção de adição (Identify in the text a causal conjunction and an additive conjunction)

A validity table for item 12 is presented below as table 49.

Table 49: Analysis of justifications for item 12 EAP teacher 2

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to recognize conjunctions for the establishment of the cohesion of the text.	This requires knowledge of metalanguage – conjunctions.	Metalanguage has probably been used during the EAP classes.

Is the item defensible? Yes. Recognizing conjunctions assists the process of establishing cohesion which contributes to the processing of relating and integrating information and developing arguments. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, the ability to recognize conjunctions and establish cohesion.

Item 13

Item 13 is as follows:

Retire do texto quatro grupos nominais com suas respectivas paráfrases⁴⁰ (Identify in the text four nominal groups with their respective paraphrases).

⁴⁰ Same as item 11.

A validity table for item 13 is presented below as table 50.

Table 50: Analysis of justifications for item 13 EAP teacher 2

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that test taker is able to recognize nominal groups with the heads and modifiers, showing ability to parse sentences.	This requires knowledge of metalanguage – nominal groups.	Metalanguage has probably been used during the EAP classes.

Is the item defensible? Yes. Since this is a situation in which the two languages – the native and the target – differ, this type of item is especially relevant in assessing the reading ability of Brazilians, in that it aims at assessing the lower level processing of sentence parsing. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, recognition of nominal groups with the heads and modifiers, and the ability of parsing sentences.

Item 14

Item 14 is as follows:

Diga a que se refere os seguintes referentes contextuais (What do the following referents refer to?)

A validity table for item 14 is presented below as table 51.

Table 51: Analysis of justifications for item 14 EAP teacher 2

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence that reader is able to establish pronoun and lexical cohesion.	This requires knowledge of metalanguage – suffixes and prefixes.	Metalanguage has probably been used during the EAP classes.

Is the item defensible? Yes. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, establishment of lexical cohesion, which contributes to local comprehension and development of micropropositions.

Item 15

Item 15 is as follows:

Traduza o segundo paragrafo do texto (Translate the second paragraph of the text)

A validity table for item 15 is presented below as table 52.

Table 52: Analysis of justifications for item 15 EAP teacher 2

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to read for understanding plain meaning.	It requires skills which go beyond the construct of reading.	

Is the item defensible? No. Although the task of translating may be part of the syllabus of the EAP courses (in terms of its objectives for preparing students for the proficiency test required for graduate students), making the test an achievement test, the skills involved in translation are different from the skills involved in the ability of reading, and are not part of the construct of reading ability used in this research. This item seems to have stronger argument against its use to provide evidence converging to the inference of reading ability.

General conclusions – EAP teacher 2

The items analyzed show that it is mostly a monomethod, multiskill, multilevel test. It is generally construct representative. Item 4 is a good representative of the construct, since it elicits evidence of the ability of the reader to integrate information around the person whose personality is described in the text, to make inference about the meaning of a word referring to his personality trait.

Item 9 is also a good representative of the construct. Although it requires knowledge inferencing, the knowledge required refers to the pragmatic knowledge in terms of functional knowledge, as described in Bachman and Palmer (1996), rather than scriptal knowledge, as described by Pearson and Johnson (1978).

There is however, construct irrelevance. Item 2 can be considered a passage-independent item, hence not assessing reading ability. Item 6 can only be answered with the knowledge of the words given in isolation as the options, whose meaning cannot be inferred with the help of context, therefore assessing vocabulary knowledge. Also, one problem for reliability is item 8. It may have low discrimination value since test takers with the ability may disagree with the key.

In the next chapter, I present the analysis of the University Entrance Examinations.

CHAPTER VI

Investigating the Defensibility of Items within the University Entrance

Examinations: Evidential Basis and Consequential Basis

In this chapter, I present the analysis of the university entrance examinations. In section 6.1, I introduce the evidential analysis of the 2003 UFSC entrance examination. In section 6.2, I present the evidential analysis of the 1998 UNICAMP entrance examination, and the analysis of some items used in examinations in subsequent years to illustrate points within a validation study. Both analyses also involve considerations of the criterion in terms of domain of reference as the tasks required for university studies. These analyses will provide information to judge the defensibility of the items for the interpretation of reading ability and for the utility of the test, to predict future performance in the criterion.

Thus, in both cases, I first present the analysis based on the framework characteristics of the test tasks proposed by Bachman (1990), and Bachman and Palmer (1996), as explained in chapters 2 and 4, to evaluate the correspondence of test task characteristics and the TLU situation characteristics. This will provide criterion-related evidence for the validation task and to judge the appropriateness of the use of the test items. Second, I bring the evidential analysis with the validity table. In section 6.3, I present an appraisal of the consequences of the use of the items for both UFSC examination and UNICAMP examination.

6.1 – UFSC entrance examination: Evidencial Basis

The examinations used for entering UFSC are developed by an institute called COPERVE (Comissão Permanente do Vestibular), which has made available the actual tests, the specifications and also a Review Book⁴¹ of past exams. This review book presents the correct answers to be given for each item and comments on the candidates' performance in terms of statistical analysis. It also provides the percentage of correct answers, the facility/difficulty degree predicted and the degree obtained for each item, in terms of a three-band scale – easy, medium, difficult.

The examinations are analyzed considering the criterion in terms of the tasks required for university studies, as well as following the constructs proposed in this research for construct validation. The analysis provides information to judge the defensibility of the items for the adequate interpretation and appropriate use of the test.

The examinations analyzed were those used for entrance in 1998, 1999, 2000, 2001, 2002 and 2003. They are very similar in terms of texts and items. They all have texts within the same topics, 12 items using the same method for assessing different skills. So it is a monomethod examination, but multilevel in that they assess comprehension both at global and local levels, and multiskills in that they assess different reading skills. Since they are rather similar, only the analysis of the 2003 examination is presented in this research. This selection was made randomly.

⁴¹ I have called *Review Book* the publication which is called in Portuguese *Provas Comentadas*.

A word on the method

The items used aim at eliciting selected responses, being what Popham (1981) calls a set of binary items, since each item is made up of a command/stem/question, followed by a set of propositions or phrases, usually from five to seven propositions, each one being either true or false.

Analysis of the characteristics of the test tasks

Bachman's (1990) and Bachman and Palmer's (1996) framework for the analysis of test content and for the analysis of the correspondence between the characteristics of TLU situation tasks (criterion) and test tasks is used here. Since the examinations used for the different years have mostly the same characteristics, the analysis reported here may be generalized to examinations used in different years.

The correspondence is judged in terms of low, medium, or high, based on the researcher's analysis of the task characteristics of the examinations, as compared to the researcher's analysis of the task characteristics of the criterion discussed in chapter 4. A high correspondence should be expected to allow generalization to performance in the TLU situation (criterion), based on the performance on the tests. For the detailed analysis, see appendix 10.

Based on the framework of the analysis of the correspondence, it is possible to see that there are many significant facets in which the correspondence is low, fact which may affect interpretation in relation to performance in the criterion, that is, fact that may not allow generalization from performance in the test to performance in the TLU situation (criterion), considering the demands of the tasks for university studies.

They are: characteristics of the setting, imposing affective schemata to be activated, length of the texts, text type, task type, pragmatic characteristics of the input (functional and pragmatic characteristics), and topical characteristics.

Some of the facets with low correspondence present greater problems for the utility of the test to predict future performance in the criterion. They are: text types, since text types determine the reading purpose, which determines the operations involved in reading (Lorch Jr., Klusewitz, & Lorch, 1995); pragmatic characteristics in terms of functional characteristics, since there are texts in the test with manipulative (regulatory) function; pragmatic characteristics in terms of the sociolinguistic characteristic of register, since there are texts in the test with informal language, which is distinct from the formal language of the academic text. This is argument to conclude that the authenticity of the test characteristics is low for some facets, and to consider the test as having low degree of utility for its purpose of selecting candidates for the university studies.

Next, analysis of an entire exam is presented concerning the 2003 examination, involving the analysis of the evidential basis within the validity table.

The 2003 Examination – validity table

The 2003 examination (appendix 11) had 4 texts, 12 test items, presented and analyzed below.

Item 1

Item 1 is as follows:

Read the summaries below. Which one(s) contains (contain) the same information found in the text?

01. According to the text, Chaplin stands as one of the greatest comedians ever, being also a relevant and powerful person in the history of the movie industry. His success is due to a character he created, known as the “little tramp”. First introduced to the world in *Kid Auto Races at Venice*, the “little tramp” appeared in all Chaplin’s movies and earned money and fame. Chaplin was meant to be successful since the beginning of his career. Born into a rich family, Chaplin was sent to an orphanage when his father and mother died. In 1912 he went on a tour with Karno’s music hall troupe, but his first performance on stage was in 1894. When touring with Karno’s group, Chaplin was invited to film *Keystone* for 150 dollars a week.

02. In the text it is said that Chaplin gained one of the highest positions as a comedian in the cinema world. The text also describes the character that brought Chaplin fame and fortune and shows when his career blossomed. Besides that, we are told about who was responsible for his recognition and the number of times Chaplin performed his “little tramp” character in the films he took part in during those years. On the other hand, we learn how difficult Chaplin’s life was when he was very young.

04. The text refers to Charlie Chaplin as one of the greatest comic actors in the whole history of motion pictures. It also tells how Chaplin gained success through the creation of his famous character – the “little tramp” – and presents a brief description of him. Besides that, the reader is informed about the hard times Chaplin had to overcome still as a child, since his father left and his mother became seriously ill. The text also mentions when Chaplin’s talent was recognized and who took part in this process. Finally, the last lines of the text show us that Chaplin played the role of his new character in many films he made at that time.

08. Chaplin, the greatest comedian in the history of motion pictures, started his career in 1914, with *Kid Auto Races at Venice*. After having lived in an orphanage, he made his first public appearance in 1894, with his mother. Chaplin was invited to work for *Keystone Films* by Mack Sennet, who brought him fame and fortune.

A validity table for item 1 is presented below as table 53.

Table 53: Analysis of justifications for item 1 UFSC entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to read carefully and summarize the ideas into the macrostructure of the text.	Summarizing into the macrostructure involves knowledge-based inferences and are, thus, rather unique.	It is a selected-response item and reader must find the most suitable sub-items to match his/her own.
	Summarizing involves understanding main ideas, distinguishing relevant from irrelevant ideas, organization of the thoughts about the text (Alderson, 2000).	The traditional summary task involves writing.	This summary task does not involve writing, only choosing either true or false for each sub-item.
		These summaries are too long, overwhelming the reader with information, affecting performance, and adding construct-irrelevant difficulty.	

Is the item defensible? Yes, with shorter summaries. This summary item allows for the assessment of high level of comprehension, as part of the integration process which involves inferences necessary for the building of a coherent meaning representation of the text, as put forth by Gagne et al (1993), and explained in chapter 3.

The fact that they are selected-response items may avoid two problems. One refers to the risk of having macrostructures unique to the individual test takers, which may differ from the test users and, consequently, be considered wrong. The other refers to the fact it avoids writing, a contaminating factor for the assessment of reading ability.

However, the summaries are too long, overwhelming the reader and resulting in fatigue, which has been considered a problem in reliability, since the assessment of the ability with the same test taker may provide different results in different testing occasions. This may be a cause of getting an item wrong for the ‘right’ reason, that is, the competent reader may be able to comprehend the text without really showing it through the item due to fatigue, thus providing low discrimination value.

With shorter summaries, this item will have only arguments in favor of its use to provide evidence converging to the inference of reading ability, in this case, being able to show global comprehension, by verifying summaries in the item representing summaries of the whole text.

Item 2

Item 2 is as follows:

Choose the proposition(s) in which the definitions of the underlined words correspond to the meaning used in the text.

- 01. figures – numbered drawings or diagrams in a book.
- 02. cane – to punish someone, especially a child, by hitting them with a long thin stick.
- 04. role – the character played by an actor in a play or film.
- 08. stage – the raised floor in a theater on which plays are performed.
- 16. dozens of – a lot of.
- 32. featured – showed.

A validity table for item 2 is presented below as table 54.

Table 54: Analysis of justifications for item 2 UFSC entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to find the appropriate meaning of a word for the context or to infer its meaning, both essential for reading.	The definitions of the target words are short.	They are enough since the test taker will add the information in the definition with the information provided within the context.
		One target lexical phrase <i>dozens of</i> is cognate. The target becomes the lexical phrase <i>a lot</i> , given as response, without any context.	<i>A lot</i> may be considered part of basic vocabulary, required for a threshold level, although not common in the criterion.

Is the item defensible? Yes. This is a type of item which tends to assess the ability to infer meaning of unknown words present in the text, without the problem of becoming a vocabulary item, which presents words in isolation, without context, as seen previously. This item has stronger arguments in favor of its use to provide evidence converging to the inference of reading ability, in this case, being able to process lexical access or infer the meaning of unknown words.

Item 3

Item 3 is as follows:

Select the proposition(s) in which the beginning of the sentence can be correctly matched with both alternatives, according to the text.

01. With his “little tramp” character Chaplin
 a) received a large amount of money.
 b) became famous all over the world.
02. Charlie Chaplin’s beginnings were not easy because
 a) his family had serious problems.
 b) his father abandoned him and his mother got a mental illness when he was just a little boy.
04. As a “little tramp” Chaplin used to wear
 a) loose trousers.
 b) a hat with a round hard top.
08. In 1912 Chaplin
 a) traveled with a music hall company around the United States.
 b) made a show with his mother.
16. Every month Chaplin
 a) received almost two hundred dollars.
 b) was invited to make a new film.

A validity table for item 3 is presented below as table 55.

Table 55: Analysis of justifications for item 3 UFSC entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to understand information related to the main theme or character, essential for comprehension. Assessing whether reader is able to identify what relates and what does not relate to the main character may provide such evidence.	The item is a little confusing to understand for the one unfamiliar with it, adding construct-irrelevant difficulty. A test taker may be able to understand the information related to the main character without being able to show it through the item.	Instructions with examples may help the understanding of the task, thus eliminating the construct-irrelevant difficulty.

Is the item defensible? Yes, as long as it is made clear and is around the main topic/theme or main character. This item seems to have stronger argument in favor of

its use to provide evidence converging to the inference of reading ability, in this case, comprehension around the main theme.

Item 4

Item 4 is as follows:

The statements in italics⁴² below were extracted or adapted from the text. They are all correct. Choose the proposition(s) in which the statement in letter a) is correctly explained or interpreted in letter b), according to the text.

01.

- a) Chaplin avoided using the new technology for some of his films but embraced it in his *The Great Dictator*.
- b) Chaplin decided to introduce sound to many films but didn't accept to use it in *The Great Dictator*.

02.

- a) The introduction of sound to the cinema brought an end to Chaplin's greatness.
- b) When silent films disappeared fame deserted Chaplin.

04.

- a) Chaplin's glory days were over.
- b) Chaplin's fame was gone.

08.

- a) "It places me on a far higher plane than any politician."
- b) The artist compares himself to a politician, and as a clown he feels less important.

16.

- a) Chaplin left the U.S. and, having been refused re-entry, made his home in Switzerland.
- b) Chaplin decided to live in Switzerland because the American people finally accepted his bad manners.

32.

- a) Chaplin's leftist politics brought him in for a good deal of criticism.
- b) Chaplin's political ideals provoked a lot of criticism against him.

A validity table for item 4 is presented below as table 56.

⁴² In this research, they are all options 'a'.

Table 56: Analysis of justifications for item 4 UFSC entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to read for local comprehension. It uses complete sentences rather than using incomplete sentences for matching.	The item is confusing for the ones unfamiliar with it, thus adding construct-irrelevant difficulty. The item may be answered without reference to the text.	Instructions with examples may help the understanding of the task, thus eliminating the construct-irrelevant difficulty All items are in English and reading them also shows reading ability.

Is the item defensible? Yes. The item focuses on local comprehension. It assesses the ability to recognize equivalent sentences, which requires elementary inferencing of reorganization or reinterpretation of information. It uses propositions rather than words. Also, the test taker will not need to produce anything, with the risk of the contaminating factor of writing. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, reading for local comprehension.

Item 5

Item 5 is as follows:

Select the proposition(s) which contains (contain) correct references to the following words, underlined in the text

- 01. which (paragraph 1): the films
- 02. their (paragraph 1): Mary Pickford, Douglas Fairbanks, D.W. Griffith, and Chaplin
- 04. its (paragraph 2): the circus clown
- 08. it (paragraph 2): the new technology

16. them (paragraph 2): several tributes
 32. the actor (paragraph 2): Chaplin
 64. that (paragraph 2): the actor

A validity table for item 5 is presented below as table 57.

Table 57: Analysis of justifications for item 5 UFSC entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to recognize and interpret cohesive devices and interpret lexical cohesion, ability essential for plain comprehension (Nuttall, 1996).		

Is the item defensible? Yes. This item assesses the ability to identify pronominal reference (subject, object, and possessive), class inclusive anaphora – *the actor* (Pearson & Johnson, 1978, p. 124), and the relative pronoun *which*. In addition to focusing on this ability of establishing cohesion of different types, very important for comprehension to occur, it does not have the disadvantage of providing construct-irrelevant easiness when options are given. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, establishment of cohesion.

Item 6

Item 6 is as follows:

Identify the correct proposition(s) according to the text

01. As Chaplin's reputation increased, so did his salary and power.

- 02. Chaplin could ask for a large amount of money for his movies after becoming famous.
- 04. For many of the films he saw, Chaplin composed the music.
- 08. Chaplin's style of performance was taken from the circus clown and mime.
- 16. After sound was introduced to the cinema, Chaplin's performance did not work its magic anymore.
- 32. Chaplin tried to re-enter the United States, but was not allowed. So he established himself in Europe.

A validity table for item 6 is presented below as table 58.

Table 58: Analysis of justifications for item 6 UFSC entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to identify main ideas resulting from more local comprehension of sentences or paragraphs.		

Is the item defensible? Yes. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, reading for local comprehension.

Item 7

Item 7 is as follows:

Select the correct proposition(s) according to the text

- 01. Among the many well-known awards given in the United States, the Nobel Prize is the most famous one.
- 02. The first Pulitzer Prizes were awarded by Joseph Pulitzer, a publisher of the New York World.
- 04. Music is one of the categories awarded by both the Pulitzer Prizes and the Grammy.
- 08. The prizes mentioned in the text were all named after outstanding people.
- 16. The name Oscar was probably a tribute to Margaret Herrick's uncle.

A validity table for item 7 is presented below as table 59.

Table 59: Analysis of justifications for item 7 UFSC entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to compare propositions and choose the correct ones, showing the ability to develop propositions based on the text.	Propositions in the items differ in small details from the ones to be developed through the text, which may escape the reader, when reading for more general comprehension.	The reading purpose is dictated by the demands of the item, which may require reader to go back and read for specific information.

Is the item defensible? Yes. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, reading for local comprehension.

Item 8

Item 8 is as follows:

In which paragraphs can you find the following information? Select the correct proposition(s) according to the text

01. The probable origin of the name of a premium given to important contributions to the film industry: paragraph 3
02. The approximate amount of categories that receive a statuette in the world of the recording industry: paragraph 4
04. The name of a country where famous rewards are delivered: paragraph 1
08. The name of a prize that is awarded monthly since the beginning of the century: paragraph 2
16. The year in which the name “Oscar” was first used to name a gold-plated statuette: paragraph 3
32. How long the person who endowed the Pulitzer Prizes lived: paragraph 2
64. The price of the gold medal that is delivered as a Pulitzer Prize: paragraph 2

A validity table for item 8 is presented below as table 60.

Table 60: Analysis of justifications for item 8 UFSC entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that the reader is able to develop propositions based on paragraphs/ develop propositional inferences and compare to the propositions given in the item.	Propositions are given in the item. Reader may just use matching.	Matching propositions will require reading ability, with the advantage of avoiding the contaminating factor of involving writing.

Is the item defensible? Yes. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, development of propositions based on longer discourse.

Item 9

Item 9 is as follows:

Which of the following questions can be answered according to the information contained in the text?

01. How much do Americans spend on awards and medals given to famous people around the world every month?
02. What is the name of the artist who received a Pulitzer Prize last year?
04. What was the first song to receive a Grammy?
08. What do people win a Pulitzer Prize for?
16. Who won an Oscar for Best Director this year?
32. How many premiums are mentioned in the text?

A validity table for item 9 is presented below as table 61.

Table 61: Analysis of justifications for item 9 UFSC entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to read for explicitly stated information, and associate the specific information with the questions given in the item.		

Is the item defensible? Yes. Reading for specific information is part of the reading ability, as well as reading for explicitly given information, both requiring literal comprehension. Items focusing on literal comprehension for reading tests are not easy to design since, in case of selected-response items, they may provide test takers with unintended cues, if information in the item (given as options) is repeated from the information explicitly given in the text. Accordingly, in case of production-responses, they will involve writing as a contaminating factor. This item avoids both problems for validation. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, reading for explicitly stated information.

Item 10

Item 10 is as follows:

The following is the last paragraph of the text. Select the proposition(s) that presents (present) the correct punctuation

01. Camera eyes are generally more accurate, than the eyes of men and women when a man looks at the world. He sees only what he chooses to see. He often finds it more convenient not to notice certain things. But, a camera represents every object completely and truthfully. Without this instrument many scientific

discoveries. Would be impossible and we would be less sure of many historical facts.

02. Camera eyes are, generally, more accurate than the eyes of men and women. When a man looks at the world, he sees only what he chooses to see. He often finds it more convenient not to notice certain things. But a camera represents every object completely and truthfully without this instrument. Many scientific discoveries. Would be impossible and we would be less sure of many historical facts?

04. Camera eyes are generally more accurate than the eyes of men and women. When a man looks at the world, he sees only what he chooses to see? He often finds it more convenient not to notice certain things; but a camera represents every object completely and truthfully. Without, this instrument many scientific discoveries would be impossible! And we would be less sure of many historical facts.

08. Camera eyes are generally more accurate than the eyes of men and women. When a man looks at the world, he sees only what he chooses to see. He often finds it more convenient not to notice certain things. But a camera represents every object completely and truthfully. Without this instrument, many scientific discoveries would be impossible and we would be less sure of many historical facts.

16. Camera eyes are generally more accurate than the eyes of men and women. When a man looks at the world he sees only. What he chooses to see? He often finds it more convenient not to notice certain things. But a camera represents every object completely and truthfully. Without this instrument, many scientific discoveries would be impossible and we would be less sure of many historical facts!

A validity table for item 10 is presented below as table 62.

Table 62: Analysis of justifications for item 10 UFSC entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to establish the syntactical relations of the sentences correctly and use correct punctuation, showing grammatical knowledge.		

Is the item defensible? Yes. Answering this item correctly requires knowledge of the syntax of the language and the ability to parse (establishment of syntactical relations) sentences correctly. This adds to construct validity since correct parsing will

help the construction of the text model. Although the correct parsing/syntax may also be assisted by the information coming from context, hence from higher-order operations, it will provide the basic evidence for the interpretation of the sentences, paragraphs, and discourse as a whole. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, the parsing of the text.

Item 11

Item 11 is as follows:

Select the correct proposition(s) to complete the following sentence: the text makes reference to...

- 01. Travels around the Americas.
- 02. The contribution of movies and photographs to our knowledge of the world.
- 04. The fact that movies and photographs can help us learn.
- 08. The stories of famous people.
- 16. The habits of rich people.
- 32. Historical American events.
- 64. An easy way to learn about other countries

A validity table for item 11 is presented below as table 63.

Table 63: Analysis of justifications for item 11 UFSC entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to identify main topics of the text.		

Is the item defensible? Yes. In addition to assessing the ability to identify the topics of the text, it also assesses the ability to use elementary inferencing. For

example, in the item we have *the contribution of movies and photographs to our knowledge of the world*, whereas in the text we have the idea that watching movies helps discovering what happens in other parts of the world. The word *help* is reinterpreted into *contribute*. The phrase *what happens in other parts of the world* is reinterpreted into *our knowledge of the world*. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, the ability to identify topics of the text.

Item 12

Item 12 is as follows:

Which proposition(s) shows(show) the main idea of all texts, according to their sequence?

01.

Text 1 – Chaplin’s beginnings and how he achieved success.

Text 2 – Chaplin’s glory, how he lost his fame and what happened in his life until he died.

Text 3 – Awards and medals that people receive all over the world.

Text 4 – The importance of movies and photographs.

02.

Text 1 – Chaplin’s life.

Text 2 – The decline of silent films and Chaplin’s death.

Text 3 – The Nobel Prize – one of the most important awards.

Text 4 – The importance of photographs in representing knowledge.

04.

Text 1 – The positive responses of cinema audiences to Chaplin’s new character.

Text 2 – The tributes received by Chaplin close to the end of his life.

Text 3 – People’s opinion about the different rewards for talents.

Text 4 – The facility of learning about other countries.

08.

Text 1 – An account of Chaplin’s career and some other biographical notes about him.

Text 2 – Chaplin’s fame and decline and what happened to him up to his death.

Text 3 – Premiums given to people in different fields of activity.

Text 4 – Movies and photographs in our lives.

16.

Text 1 – A description of Chaplin’s most important character.

Text 2 – Chaplin’s death.

Text 3 – The origin of some of the very well-known statuettes awarded every year.

Text 4 – The autonomy man has in choosing what he wants to see

A validity table for item 12 is presented below as table 64.

Table 64: Analysis of justifications for item 12 UFSC entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to develop the macroproposition of each text.		

Is the item defensible? Yes. The item requires the recognition of the macroproposition of the text, i.e., the proposition(s) reflecting the main idea of a text. Answering this item correctly requires the knowledge of the syntax of the language and the ability to parse sentences correctly, to recognize the words and access their appropriate meanings, and go through the higher-level processes of inferential comprehension, both integration and summarization. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, the ability to develop macropropositions of texts.

General conclusions concerning evidential basis

This examination may be mostly considered a comprehension test, with focus on comprehension questions. This can be concluded based on the number of items focusing on the recognition of propositions, summaries, topic, and macropropositions.

It also has items focusing on micro-skills, for example, on words to be lexicalized in context such as pronouns or text-structuring words for lexical cohesion (Nuttall, 1996).

Nevertheless, some problems may be pointed out. There may be construct-irrelevant difficulty in terms of unfamiliarity and complexity of some items. There may also be construct-irrelevant easiness in terms of the topics presented, since they all belong in the same subject of cinema, movies and photography.

Another problem refers to reliability. There is a general orientation that any assessment instrument have, at least, 12 items for securing reliability. However, some of the summaries are too long and with too many details, which, together with many sub-items, sometimes 6 or 7, may cause fatigue. This is a problem for reliability, since results in a fatigue situation will not be consistent with results in a non-fatigue situation.

A major problem for the examination is that the assessment of the authenticity based on Bachman (1990) and Bachman and Palmer's (1996) framework showed low degree of correspondence with some facets, particularly low with some of text types used in the examination when compared to the texts used in the criterion (found with the use of the framework for authenticity). Since text types determine skills adopted for reading, it can be concluded that the examinations may not have the expected predictive power to predict performance in the future, in the criterion.

6.2 – UNICAMP Entrance Examination: Evidential basis

The examinations used for entering UNICAMP are developed by an institute called COMVEST (Comissão Permanente para os Vestibulares), which has made available not only the criteria and the methods used, but also a Review Book for each

examination. This Review Book presents the expected answers to be given for each item with comments as to why they are expected, and a discussion of candidates' performance in each test item. It also presents comments on the candidates' performance in terms of statistical analysis within their six-band scale – zero to five, where zero corresponds to nothing acceptable and five corresponds to the expected answer.

The examinations are analyzed considering the criterion in terms of the tasks required for university studies, as well as following the constructs proposed in this research for construct validation. The analysis will provide information to judge the defensibility of the items for the adequate interpretation and appropriate use of the test.

The examinations analyzed were those used for entrance in 1998, 1999, 2000, 2001, and 2002. They are very similar in terms of texts and items. They all have 12 items with the same method assessing different skills. So it is a monomethod examination, but multilevel in that the items assess comprehension both at global and local levels, and multiskills in that they assess different reading skills. Since all the examinations are very similar, out of the five analyzed, the 1998 examination was chosen randomly, and the analysis of the whole examination is presented in this study. In addition to that, the analysis of some items of the examinations used for the entrance in other years is also presented.

A word on the method

The examination uses only items, rather than prompts. The case here is limited-production⁴³ in Bachman and Palmer's (1996) terms or open-ended questions as used by most researchers in reading.

Open-ended questions allow respondents to, in the process of reading and understanding a text, use their ability of inference and apply their previous knowledge. They have the advantage of allowing test takers to think for themselves, without having indication of the right answer (Nuttall, 1996, p. 186). However, since respondents' knowledge is involved in the process of reading, resulting in variation in comprehension (Urquhart & Weir, 1998), since different readers arrive at different understanding of a text (Alderson, 1996, p. 225), since the resulting mental representation is a combination of text model and situation model (Kintsch, 1998) as discussed in chapter 3, resulting in the understanding of the text as well as the interpretation of the text, unique to every reader (Grabe, 1999), it is difficult to determine what is the 'correct understanding'⁴⁴.

This kind of question should not be expected to elicit one correct answer, and all possible or plausible answers must be considered by the test raters/users. Pearson and Johnson (1978) talk about plausible answers – those which allow the generation of an argument in support of their plausibility, and textual intrusions – those coming from the text, but not allowing the generation of an argument in support of their plausibility. However, Norton and Stein (1998) talk about divergent comprehension, claiming that

⁴³ The terminology used here follows Bachman and Palmer (1996) as defined previously. Some scholars, such as Scaramucci (2002), call the type of questions used in this exam as open-ended.

⁴⁴ The expected answers for the entrance examination reflecting the correct understanding are made public within their Review Book available on the Internet.

there are some interpretations which are also legitimate. This raises the methodological and the ethical problems, discussed in chapter 3, of who has the ‘correct understanding’ of the texts to judge the answers given by the test takers as right or as wrong? Who, or which group, is the ‘correct understanding’ normed on?

There is another problem with the use of open-ended questions: it involves the ability of writing, which is considered a contaminating factor for the assessment of reading, since reading is a receptive skill, with specific operations, and writing is a productive skill with different operations. Using writing means using a construct other than reading, and is considered a source of invalidity.

Analysis of the characteristics of the test tasks

As with the previous analysis of UFSC entrance examination, Bachman’s (1990) and Bachman and Palmer’s (1996) framework for the analysis of test content and for the analysis of the correspondence between the characteristics of TLU situation tasks and test tasks is used for the analysis of this university entrance examination. Again, the correspondence is judged in terms of low, medium, or high, and a high correspondence should be expected to allow generalization to performance in the TLU situation based on the performance on the tests. The analysis in this study may be generalized to the other examinations, in that they have mostly the same characteristics. For the detailed analysis, see appendix 12.

Based on the framework of the analysis of the correspondence, it is possible to come to a very similar conclusion to the previous analysis about UFSC entrance examination analyzed, with respect to the low correspondence. However, in this case, the low correspondence is found in more facets, namely, in the facet pragmatic characteristics, in terms of functional characteristics, with texts in the test with

manipulative function (advertisements) and imaginative function (poetry and comic strips) and pragmatic characteristics in terms of sociolinguistic characteristic of register, with informal language, hence very distinct from the formal language of the academic text.

The low correspondence, thus, refers mostly to the task characteristics which, in the criterion, are determined by the use of academic texts, but in the test, are determined by texts such as ads, poems, comic strips, narratives, with figurative language such as proverbs, idiomatic expressions, and tone of irony, for example. It is relevant to mention that the task characteristics determined by these types of texts already cause differences in performance, but more importantly for the analysis in this study is that the items of the examinations focus on these characteristics with low correspondence.

The conclusion for this examination is that the inferences concerning language ability based on the performance on the test cannot be generalized to language ability in the criterion, i.e., what is concluded about the performance of a test taker cannot predict his/her future performance in the criterion. Thus, action inferences – allowing or denying admission in the criterion (university studies), is not supported by the interpretation based on the test score.

Further analyses to illustrate not only the low correspondence but also items focusing on the low correspondence are given below, using some items from UNICAMP examinations administered in other years.

One refers to item 9, 2000 examination (appendix 13). The reader is presented with an excerpt of a story/narrative taken from *The Victorian Fairy-Tale Book*, and with the question *explain how he [the protagonist] comes to change his mind.*

Answering this question depends on understanding of the meaning of a proverb⁴⁵, which contains the very colloquial expression *never mind it*, essential for the answer. Proverbs are culture-specific and are usually avoided in academic texts.

Item 2, 2000 examination (appendix 14), is also an example. The reader is presented with an excerpt of book, and with the question *what is the prediction by Mimi about John Lennon' future?*. Answering this question depends on understanding the idiomatic expression *make a living out of it*, and idiomatic expressions are culture-specific and usually avoided in academic texts.

Another example is item 16⁴⁶, 2001 examination (appendix 15). The reader is presented with a poem, and with the following question: *how does the poem by Carroll Arnett explain that Your problem is not my problem?*⁴⁷. Answering this question correctly requires, according to the Review Book, being able to capture the irony implicit in the text. Irony is culture-specific and is the kind of language usually avoided in academic texts.

Another example is item 14⁴⁸, 2002 examination (appendix 16). The reader is presented with a short narrative, and with the following question: *who are the characters to appear in the passage?* and *How do they relate?* Answering the second part of the question requires recognizing the point of view of the narrator by “distinguishing facts observed and reported by the narrator from facts that he [the narrator] presents as taken for granted by the character” (Review Book). This skill is typical for reading literary texts, not academic texts, and it is listed in the taxonomies of

⁴⁵ In the Review Manual, it is recognized that the difficulty in the understanding of the proverb was the main source of problems to answer the item correctly.

⁴⁶ Although it is only the fourth item in the test, it is numbered 16. I have decided to keep the original as published on the Internet.

⁴⁷ *Your problem is not my problem* is the first sentence of the poem.

⁴⁸ Although it is only the second item in the test, it is numbered 14. I have decided to keep the original as published on the internet

skills⁴⁹ for narratives, not in taxonomies of general reading or academic reading, although it may be argued that it is compared to distinguishing facts from opinions, which is in taxonomies of skills for academic purposes.

The analysis of all these items lead to the conclusion that there is low authenticity of the tasks as judged by the correspondence of task characteristics in the test and task characteristics in the criterion, the target language use situation, and that performance based on the items in the examinations cannot predict performance in the criterion/target language use situation.

Next, I present the analysis of an entire examination, followed by the analysis of some items from other examinations considered relevant. For the analysis, although I consider the use of writing required by the open-ended questions a contaminating factor for the assessment of reading competence, I will consider this irrelevant for further mention, since it is a characteristic of all the items. I will, however, resort to comments using this source of invalidity when I consider extreme cases, which is explained for each individual item.

The 1998 Examination

The 1998 examination (appendix 17) was comprised of 6 texts, 12 test items.

Item 1

Item 1 is as follows:

Quem é quem nessa história? (Who is who in this story?)

⁴⁹ Taxonomies of skills are given in appendix 19.

A validity table for item 1 is presented below as table 65.

Table 65: Analysis of justifications for item 1 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence that the reader is able to establish cohesion.		

Is the item defensible? Yes. Answering correctly to item 1 requires establishing cohesion of the various uses of the pronouns, in particular *she* and *her*. There are three characters in the story: Harold, Doll, and Rosie. Assuming that the test taker is able to recognize that Harold, Doll and Rosie are people by use of capital letters, that Harold is a man and Doll and Rosie are women, by using their knowledge of names, and that *she* and *her* are pronoun words to refer to women, the decision for the answer will be based on the two women. Establishing reference might involve two levels of operations: establishing cohesion (or interpreting pro-forms as referred to by Nuttall, 1996), just by recognizing that *she* and *her* refer to a woman, when there is only one possible referent; or in case there are two referents, drawing inferences to determine which one the pronoun refers to. In the example of the test, the paragraph giving this information for the answer is as follows:

Rosie sat at the table and ate her(1) dinner. She(2) thought her(3) mum was being stupid, although she didn't say so. Instead, she just filled her mouth with a forkful of mashed potato and stared at her plate.

It is very simple to establish reference as required by the question. The character is Rosie and the first pronoun *her* refers to *Rosie* because *Rosie* is a near referent. For the subsequent pronouns, the argument of a near referent becomes less strong, but still,

there is no one else to refer to in the whole paragraph. The reader will easily establish the cohesion between the first *she*(2) used to *Rosie*, and the second *her*(3) used to refer to *she*(2), who is ultimately *Rosie*.

In case the reader knows the meaning of *mum*, or is able to infer its meaning, or is able to recognize it as a cognate word, it will be simple to know that *Rosie* has a mother, thus she is the daughter, and since Harold is a man, by elimination, Doll would be the mother. The conclusion that Harold is the father may come from two sources: the schema of a family is activated through all the information presented such as there is a man, a mother and a daughter at a table, and the family schema would suggest that the man is the father; or the reader goes on and recognizes the word *father* used in subsequent sentences. The answer to the question is complete. This item seems to have only argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, of the establishment of cohesion.

Item 2

Item 2 is as follows:

A que se refere 'Shadow Point'? Por que recebeu esse nome? (What does "Shadow Point" refer to? Why has it received such a name?)

A validity table for item 2 is presented below as table 66.

Table 66: Analysis of justifications for item 2 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide evidence that the reader is able to relate information to interpret lexical cohesion.	It is likely to be answered by using only the picture, without reference to the text.	Information in the text will provide more evidence for confirmation.

Is the item defensible? No. Since there might be the construct-irrelevant easiness factor. Answering item 2 correctly requires understanding the words *shadow* and *point*. The item, however, will most likely be answered just by relating the question given in the native language Portuguese to the picture, without reference to the text. According to Nuttall (1996), questions to assess reading competence should not lend themselves to be answered without the reading of the text (p. 190). In fact, items that can be answered without reference to the written text will not provide any evidence of reading competence in a foreign language, thus providing construct-irrelevant evidence. This item seems to have stronger argument against its use to provide evidence converging to the inference of reading ability.

Item 3

Item 3 is as follows:

O texto menciona mudanças. Que mudanças são estas? (The text mentions changes. What changes are these?)

A validity table for item 3 is presented below as table 67.

Table 67: Analysis of justifications for item 3 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to find specific information and reinterpret or reorganize it to provide the answer, thus drawing elementary inference.	May be answered using only the picture, without reference to the text	Confirmation may be found in the text

Is the item defensible? No. Since it may be answered without reference to the text. The question itself provides the reader with a great deal of information about the content of the text, information given in Portuguese, which together with the analysis of the picture based mostly on the readers' background knowledge of the possible consequences of the construction of the first building, extremely tall, in a small city, with its shade over most houses, provides the test takers with what Popham (1981) has called unintended clues for the answers, thus allowing the answers to be given much independently of the text. This item seems to have more argument against its use to provide evidence converging to the inference of reading ability.

Both items 2 and 3 raise the issue of the role played by illustrations. Schallert (1980) claims that illustrations can contribute to the sum of the information to be conveyed, i.e, pictures and words interact to provide the necessary information. This contribution is particularly helpful, according to the author, in the case of expository texts, since the writer must use the right words to constrain the readers' interpretive and constructive processes in such a way that readers will understand the author's intents. The author stresses that the appropriate illustration, used in an expository text, is particularly effective to help the readers comprehend new concepts and new relationships among concepts.

However, illustrations must be seen as auxiliary to the written text, an accessory to the words, or adjuncts to the text, as Schallert (1980) refers to that, not a replacement of the written text. As the authors has pointed out, readers tend to take the easiest route to understanding, which may mean using the illustration and ignoring the written text. In sum, illustrations function as an aspect of construct-irrelevant easiness, or as unintended clues, both sources of invalidity since the construct being measured is language ability in English as a foreign language. If the reader is able to respond to an item by using illustration(s) only or mainly, hence ignoring the written text, he or she is just showing comprehension, not reading in English as a foreign language.

Both items 2 and 3 may be considered passage-independent items, since they may be responded mostly with reference to the illustration, without reference to the passage. Such items must be avoided if valid interpretation is being sought; and they are considered in this research as source of invalidity. They have low discrimination value, thus not being dependable for a good discrimination between test takers with the ability and those without it, since they may be answered mostly without reading the written text, without showing the ability of reading in English as a foreign language.

Item 4

Item 4 is as follows:

O primeiro parágrafo se dirige a um publico específico. Que publico e este? Justifique sua resposta. (The first paragraph addresses a specific public. What public is this? Justify your answer)

A validity table for item 4 is presented below as table 68.

Table 68: Analysis of justifications for item 4 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to interpret the use of hypothetical <i>you</i> , find specific information related to it, and draw inference.	Hypothetical reader is different from target reader, the former not listed in taxonomies of academic reading. Justifying the answer requires a level of argumentation very demanding in terms of writing ability.	

Is the item defensible? No. The hypothetical *you* (referred to as the public that the writer has in mind, in the Review Book) is not usually used in academic texts, thus not necessarily in the criterion of reference used in this study. For it to be identified, the test taker must forget the usual information used for the identification of the target reader, rather relevant for university studies, such as the media used for publication, the register, the function, the topic. In this case, the conclusion would be that the target reader is someone with access to the Internet in that it has the format of a text published on it, someone who reads science magazine, not someone hypothetical who is in a rural area in Africa and is familiar with all commodities of modern life, such as fast food restaurants (Review Book).

Also, justifying the answer, as in the item, requires a level of argumentation which is very demanding in terms of writing ability, far beyond what can be considered acceptable for the assessment of reading competence, thus involving construct-irrelevant skill.

This is another item whose focus is, admittedly, not worked upon in the secondary school (Review Book), which, together with the fact that it is not the usual inference for the identification of the target reader, might account for the level of

difficulty found represented by 70% of score zero (Review Book). If not worked upon previously, and not in the criterion, why to use this kind of question?. What is being determined through the use of questions such as these? This item seems to have stronger argument against its use to provide evidence converging to the inference of reading ability.

Item 5

Item 5 is as follows:

Qual é a explicação de Abrahams e Pearson para o uso de adjetivos como "eccentric", "perverted", "odd" and "bizarre" para caracterizar a geofagia? (What is the explanation, by Abrahams and Parsons, for the use of adjectives such as "eccentric", "perverted", "odd" and "bizarre" for the characterization of geophagy?)

A validity table for item 5 is presented below as table 69.

Table 69: Analysis of justifications for item 5 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to relate information scattered through one paragraph.		

Is the item defensible? Yes. The item seems to be assessing the ability to identify views on the theme, requiring elementary inferencing in terms of reinterpretation and/or reorganization of information explicitly given in two subsequent sentences presented in the same paragraph, and associating synonymous information all related to the same theme. This item seems to have stronger argument in favor of its use to

provide evidence converging to the inference of reading ability, in this case, relating information for the integration process.

Item 6

Item 6 is as follows:

Dê um significado para a palavra 'but' no trecho "on the whole [soil eaters] are regarded as quite normal to most but outsiders" (Give a meaning to the word "but" in the phrase "...on the whole [soil eaters] are regarded as quite normal to most but outsiders").

A validity table for item 6 is presented below as table 70.

Table 70: Analysis of justifications for item 6 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to find the appropriate meaning for the word <i>but</i> as used in its context.	Poor context hindering inference. Occurrence of an ellipsis hindering inference.	Semantic constraint is fairly strong. Ellipsis is part of the normal reading.

Is the item defensible? No. This item seems to assess the ability to find the appropriate meaning for a word either based on lexical access, or based on inference of its meaning through the use of the context.

It may, however, be considered a trick question since *but* is most likely to be a familiar word for most readers with its meaning as a conjunction, not as a preposition as required in the item. In this case, finding its appropriate meaning depends on understanding the following word, *outsiders*, which is not frequent, and likely to be

unknown. Also, there is a case of ellipsis right before the word *but*, making it more difficult for inference of its meaning. This may account for the 66% of score zero obtained by test takers (Review Book). The 22% of score 5 may probably be accounted for as lexical access, i.e., choosing the appropriate meaning when already known, rather than inference of unknown meaning, as expected by the test raters (Review Book).

Although the target skill of meaning inference or the skill of lexical access are relevant to the interpretation of reading ability, this item seems to have stronger argument against its use to provide evidence converging to the inference of reading ability.

Item 7

Item 7 is as follows:

De que maneiras a violência urbana pode estar afetando a saúde de pessoas idosas? (In what ways can urban violence be affecting the health of the elderly?)

A validity table for item 7 is presented below as table 71.

Table 71: Analysis of justifications for item 7 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to read for global comprehension.	Unintended clues are given through the question in Portuguese. Topic is biased in favor of a group.	This may be compensated by using other topics for the other texts within the examination.

Is the item defensible? Yes. This item allows the assessment of reading for global comprehension, for the main idea and its supporting arguments. It, however, may provide unintended clues with the information in Portuguese given in the question, adding to construct-irrelevant easiness which, together with background knowledge on the topic, might help develop strong hypotheses about the text content even before reading it. This may account for the rating of the question as easy by the test raters, although the task demand is not simple.

Since it was considered the easiest among the prospective medical students, this item faces the ethical problem of being biased in favor of a group, consequently against the others in the case of a selection process. Providing texts in various areas is a way of compensating for this bias and of providing the same chances for the test takers more familiar with other areas. In so doing, validity of the interpretation is enhanced, since reading competence, not background knowledge, is what is being assessed.

Although there might be construct-irrelevant easiness, reading the text confirms the hypotheses, allowing for the expected answer to be given. This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, reading for global comprehension.

Item 8

Item 8 is as follows:

O que deu origem a estes dois textos? (What was the origin of the two texts)

A validity table for item 8 is presented below as table 72.

Table 72: Analysis of justifications for item 8 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to establish the context for the production of the text.	Establishing context may require pragmatic inferencing.	The context is explicitly given in the first text, referring to the previous publication.

Is the item defensible? Yes. Test taker may use only elementary inferences for the answer, since the information required is provided in the text. Textual schemata may help, since the texts within the section *Letters to the Editor* are usually published in response to previous articles. This may account for the 42% of scores 4 and 5, and the low occurrence of score zero, 32% (Review Book). This item seems to have stronger argument in favor of its use to provide evidence converging to the inference of reading ability, in this case, relating information given in the texts.

Item 9

Item 9 is as follows:

O primeiro texto destaca dois pontos positivos e faz uma ressalva. Transcreva o quadro abaixo para o seu caderno de respostas, preenchendo-o com as informações necessárias (The first text highlights two positive points and a limitation. Fill out the following chart with the necessary information)

A validity table for item 9 is presented below as table 73.

Table 73: Analysis of justifications for item 9 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to find specific information as positive aspects and a limitation.		

The task requires deciding on what the positive aspects and the limitation are, involving inference making on the part of the reader, since this is not explicit in the text. However, it has technical problems, analyzed below as table 74.

Table 74: Analysis of the technical quality of item 9 UNICAMP entrance examination

Technical quality	Comments
Appropriateness of the question	The question is not clear as to what the positive aspects and the limitation refer to.

Is the item defensible? Only if it can be made clear. Since the question is not clear as to what the positive aspects and the limitation refer to, this must be inferred by the reader. Although the Review Book claims that the expected answer should contain information about the positive aspects and limitation of the previous article, it is possible to argue that the mentioned law (Murphy's law) has the positive aspects required by the question, as made clear in the text: it helps 1) determining the likely causes of failure in advance, and helps 2) in the decisions to prevent the problems. This is confusing for test takers. To make things worse for the reader, this text has the following unusual confounding characteristic: the writer of the article, included in the examination, refers to the writer of the previous articles in the second person, using the pronoun *you* (4 times) and the pronoun *your* in the first paragraph, but in the second

paragraph, using the lexical item *the author*, third person, as if they were two different people. Assuming that this is a characteristic infrequent in academic texts, this adds criterion-irrelevant difficulty to the item. Considering the technical problem, this item seems to have stronger argument against its use to provide evidence converging to the inference of reading ability.

Item 10

Item 10 is as follows:

O Segundo texto afirma: “the condition of this issue is an excellent example for her presentation”. Explique por que. (The second text says: “the condition of this issue is an excellent example for her presentation”. Explain why).

A validity table for item 10 is presented below as table 75.

Table 75: Analysis of justifications for item 10 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to establish lexical cohesion, use coherence, and integrate information within and across texts.	Explaining, as in the item, requires a level of argumentation very demanding in terms of the writing ability, far beyond what can be considered acceptable, thus involving construct-irrelevant skill. This question is likely to favor those who are familiar or more familiar with the mentioned law, thus being biased.	A small piloting would determine how much knowledge of the topic is necessary for the item.

Is the item defensible? No, not the way it is designed. This is a very good item to make some points concerning reading ability. The question is: what is being assessed through this item: is it reading ability or some other construct represented by the task *explain*, as required by the item? I want to argue that it is assessing a construct other than reading ability. Considering the sentence from the text, which is the basis to answer the item *the condition of this issue is an excellent example for her presentation*, it is possible to see that it has four words - *condition*, *issue*, *example*, and *presentation*⁵⁰ - to be lexicalized through lexical cohesion or the establishment of coherence. Establishing cohesion and/or coherence for all the four words is essential for the integration of the information, thus for comprehension to occur. An item based on the task of integrating information would show the test takers' ability with these essential processes for comprehension.

Now, what is possible to interpret about the reading ability of 70% of the test takers who either gave no answer or gave the wrong answer (53% of score zero and 17% of no answer), and about the other 20% of the test takers with low scores from 1 to 3⁵¹, making up 90 % of all the test takers? Is it possible to interpret that they were not able to integrate the information or that they were not able to explain that? Explaining, as in the item, requires developing articulated reasoning/thinking and expressing it through writing. This is far beyond integrating information for comprehension, or any skill or operation involved in the construct of reading. This is a source of invalidity.

In addition to that, the task may have been too demanding for those who did not have previous knowledge about the mentioned law. In case the question requires information on the law, which is presupposed for the reader, then the question is also

⁵⁰ *Issue* refers to the specific edition, *condition* to the incorrect assembling and delivery damage, *example* to the fact that such a thing has happened, and *presentation* to the information presented previously in the text, about the daughter's 'talk' in her science class.

⁵¹ The scoring system used for these examinations involves a band ranging from 0 to 5.

assessing knowledge on the topic, hence, being a source of invalidity. This item seems to have stronger argument against its use to provide evidence converging to the inference of reading ability.

Item 11

Item is as follows:

Explique por que Murphy pode ser considerado um perfeccionista. (Explain why Murphy may be considered a perfectionist).

A validity table for item 11 is presented below as table 76.

Table 76: Analysis of justifications for item 11 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to find the subordinate pieces of information that would be integrated to the proposition of the title <i>Murphy was a perfectionist</i> , involving inferences based on association with the superordinate word <i>perfectionist</i> .	Explaining, as in the item, requires a level of argumentation very demanding in terms of the writing ability, far beyond what can be considered acceptable, thus involving construct-irrelevant skill.	

Is the item defensible? No. Although the focus on the title should provide evidence for valid interpretation, since it tends to reflect the macrostructure of a text, explaining it in writing seems to require a level of argumentation very demanding, far beyond what can be considered acceptable for the assessment of reading ability, thus involving construct-irrelevant task of explaining in writing. This item seems to have

stronger argument against its use to provide evidence converging to the inference of reading ability.

Item 12

Item 12 is as follows:

Explique o titulo do texto (Explain the title of the text).

A validity table for item 12 is presented below as table 77.

Table 77: Analysis of justifications for item 12 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to find the subordinate information related to the proposition of the title, involving inferences.	Understanding of the text might be difficult for most readers since it has many topic-specific words. Explaining demands operations far beyond the comprehension skills, involving writing skills.	The topic and the topic-related words are uncommon to all readers, thus not likely to cause bias.

Is the item defensible? No. Same as item 11. This item seems to have arguments against its use to provide evidence converging to the inference of reading ability.

Final remarks concerning the 1998 UNICAMP examination

It is possible to conclude, based on the analysis, that there are many sources of invalidity, in particular, in the items which require the task of explaining in writing in a

test designed to assess reading ability. Evidence collected through these items cannot be used for inferences of reading ability.

Also there are some sources of invalidity in terms of bias, when the items developed tend to favor one group to the disadvantage of the other(s). In this respect, the choice for the text *The soil-eaters*, may be considered adequate for testing, since it can be argued that its topic - geophagy - is probably equally unknown to test takers. According to Urquhardt and Weir (1998), traditional testing requires minimizing the effect of background knowledge, which may be achieved by choosing texts whose topic is equally unknown to all the test takers. This may, in the end, enhance validity of the interpretation of reading ability, since what is being assessed is reading as the result of the work done on the specific text, not background knowledge.

In the specifications in the Candidate's Manual, the criterion is defined as the abilities essential for university students to carry out their studies. The items examined, however, do not reflect that, and this makes the examination inappropriate as a university entrance examination. Next, I present the same kind of item analysis with some items from other UNICAMP examinations, to further the validation investigation.

1999 examination

Item 10 appendix 18.

Item 10⁵² is as follows:

Qual era o problema do Sr. Newton (What was Mr. Newton's problem?)

A validity table for item 10, 1999 examination is presented below as table 78.

⁵² In the Review Book, it is question number 22.

Table 78: Analysis of justifications for item 10, 1999 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to read for global comprehension.	It requires pragmatic inferencing for the establishment of coherence. Pragmatic inferencing requires world knowledge that may not be shared by test takers.	

Is the item defensible? No. This level of inference is implied by what Mr. Newton said, that now that the power was on, he could turn out the lights. The coherence is established by the reader when bringing to bear the knowledge that a 93-year old man may have rituals to follow before being able to sleep, rituals which may include turning the lights out. This presupposes the knowledge about old people, that they follow rituals. In case this knowledge is not shared by some test takers, which is plausible since it is not the case that every elderly follows rituals, they would not be able to answer the question. Knowledge cannot be presupposed for a language test. Clapham (1996) has shown that it is very difficult to know what the background knowledge of university students consists of.

Although this level of inference involving pragmatic inferencing is essential for comprehension, it can be argued that the item in the test has construct irrelevance, since it requires specific knowledge not recoverable in the text, a factor not part of the construct used in this research. It can also be argued that it has criterion irrelevance, since academic texts tend to be written in a way to provide the target reader with the most information essential for comprehension to occur.

What is the meaning of the low scores of zero, 1 or 2? What inferences can be made based on them? Is it possible to infer that the test taker cannot process at the lower level, to be able to understand the plain meaning of the text? This is one possibility. Is it possible to infer that the test taker is not able to make inferences? Not likely, because inferences are essential for any communication. Claiming that a person is not able to make inferences is the same as claiming the person cannot understand simple dialogs for everyday communication.

Getting the item wrong may, therefore, mean only that the test taker did not share the presupposed knowledge that old people have rituals, either not having it, or having different knowledge such as the idea that old people, nowadays, like to lead a busy life. This item seems to have stronger argument against its use to provide evidence converging to the inference of reading ability. I turn now to the analysis of the technical quality of the item. A technical quality table for the item is presented below as table 79.

Table 79: Analysis of technical quality for item 10, 1999 UNICAMP entrance examination

Technical quality	Comments
Appropriateness of the question	<p>The question is not clear as to the level of inference it requires. <i>What was Mr. Newton's problem?</i> may get as an answer that his power was out. Collins Cobuild dictionary defines <i>problem</i> as <i>a situation causing difficulties for people</i>. Having the power out may be argued to be the situation causing difficulty for Mr. Newton.</p> <p>In the Review Book, when presenting the arguments for the expected answer and commenting the level required, a different question is used: <i>what was, after all, Mr. Newton's problem?</i> It is a different question, indicating that there is another problem to be considered for the answer.</p>

Is the item defensible? No. Test takers might get the item wrong for the 'right' reason (having the ability). This may explain the 37,6 % of scores zero for this question (Review Book). Why not carry out an analysis of internal reliability? This involves

comparing the performance of test takers in different items of the same test. This might lead to the finding that many test takers in the top group got this item wrong, and to the conclusion that there is something wrong with the item, not with the test takers.

2000 examination

Item 9 appendix 13

Item 9 is as follows:

Explique como ele [o protagonista] chega a mudar de ideia (Explain how he [the protagonist] changes his mind)

A validity table for item 9, 2000 examination is presented below as table 80.

Table 80: Analysis of justifications for item 9, 2000 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to read for specific information.	Understanding specific information is based on the figurative language presented as a proverb in English, and proverbs are culture-specific, difficult to understand for those outside it.	Inference making may help understand the proverb the same way as words or expressions when the context is rich.

Is the item defensible? No. Answering the item is based on the figurative language presented as the proverb for every evil under the sun, there is a remedy, or there is none; if there is one, try to find it – if there isn't, never mind it. Although inference making may help understand the proverb the same way as words or expressions, the context is not rich enough for the inference to be made. This may account for the fact that this item was considered one of the most difficult items in the

test, difficulty attributed to the understanding of the proverb (Review book). This item seems to have stronger argument against its use to provide evidence converging to the inference of reading ability.

2001 examination

Item 16 appendix 15

Item 16 is as follows:

Como o poema de Carroll Arnett justifica que Your problem is not my problem? (How does the poem by Carroll Arnett justify that Your problem is not my problem?).

A validity table for item 16, 2001 examination is presented below as table 81.

Table 81: Analysis of justifications for item 16, 2001 UNICAMP entrance examination

Outcome: interpretation of the item as measure of reading ability in L2			
Justifications	Argues in favor	Argues against	Refutation
Evidence: Item and task analyses	Item appears to provide the evidence that reader is able to establish coherence based on the title <i>Next</i> .	Coherence, in the case, can only be established resorting to knowledge outside the text, or scriptal data, or the use of pragmatic inference	

Is the item defensible? No. Being able to reconstruct the chain of reasoning of the writer presupposes the knowledge about the inefficiency of services, about people waiting in lines to be helped by the clerk sitting behind a table (Review Book). Not being able to reconstruct the chain of reasoning of the writer to answer the question may mean either that the reader did not share that knowledge, or that *Next*, in the title, did not activate it.

The interpretation will be, however, about ability, that the reader did not have the reading ability, not that he/she did not have the knowledge, or that the knowledge was not activated. Nuttall (1996) claims that poetry is an extreme case of difficulty for the reconstruction of the chain of reasoning of the writer, being the most difficult situation to “make coherent sense of the text” (p. 103). This item seems to have stronger argument against its use to provide evidence converging to the inference of reading ability.

In sum, a common characteristic of item 10, 1999 examination; item 9, 2000 examination; and item 16, 2001 examination is that they are based on pragmatic knowledge as used by Urquhart and Weir (1998) and Hughes (2003), or on scriptal data or knowledge, as used by Pearson and Johnson (1978).

This becomes a problem for construct validation since the inference based on the test items will be about language ability, that is, if the reader does not know about old people’s rituals or has an alternative knowledge about old people, if the reader does not know the culture in which the proverb makes sense, if the reader does not know the culture in which the irony related to the public administration service makes sense, the test taker will not be able to answer the question and the inference based on the scores will be that the reader does not have the language ability or the reading ability.

It may be argued, however, that expressions, figurative language, irony, or culture-specific information may be present in all forms of communication. Although this raises the counterargument that they are infrequent in the criterion, the point here is not that this culture-specific information be avoided, but rather, that items developed to provide evidence for the inference of reading ability should not focus on these pieces of culture-specific information.

Anyway, what inference about reading ability is possible based on an item focusing on the proverb “*for every evil under the sun, there is a remedy, or there’s none; if there is one, try to find it- if there isn’t, never mind it*”? Or what inference about reading ability is possible based on the only item for a text focusing on the expression *make a living out of it*? Or what inference about reading ability is possible based on the only item for a text focusing the understanding of the irony implicit in the text to be captured by accepting that the title word *Next* refers to the next one in a long line waiting to be helped?

The only valid inference based on the answers to those questions, or rather, based on the failure to answer the questions, is that the test taker does not know the specific proverb or is not familiar with proverbs, does not know the expression, or does not share the knowledge of lines as reflecting a neglectful treatment in organizations. But little evidence for the assessment of reading ability.

General conclusions concerning evidential basis

In the candidate’s manual for the 2003 entrance examination, the then Dean of UNICAMP stresses that their entrance examination has had, since 1987, the same aims, with the use of open-ended questions, of selecting students who can think, draw correlations, develop hypotheses. This is confirmed by Scaramucci (2002), when she mentions the aims of UNICAMP entrance examinations as “higher-order cognitive skills such as the ability to organize and express ideas clearly, to establish relationships and interpret results, formulate problems and develop hypotheses” (p. 64). In fact, based on the analysis carried out in this research, it is possible to come to similar conclusions as to the aims of the items.

However, in the same candidate's manual, it is stated clearly, on page 37, that the objective of the foreign language examination is to assess reading competence, information which is repeated on page 38. Based on the construct of reading ability, it can be argued that expressing ideas clearly is a source of invalidity, since it is only possible through the productive skills of speaking or writing. In fact, writing is required extensively by the items aimed at assessing reading competence, which is a source of invalidity in itself.

It is also clearly stated in the manual that the examinations will assess reading competence since "reading in a foreign language is very useful for the university students to carry out their studies" (p. 37)(my translation). The criterion or target language use situation is, thus, defined, being the demand for carrying out the university studies. This does not seem to be what was concluded based on the analysis of the items, and based on the analysis of the test content with the use of Bachman's (1990) and Bachman and Palmer's (1996) framework. Aside from specific cases, classes in the criterion do not usually require reading poems, poetry, narratives, comic strips, proverbs, ads, novels, letters to the editor, fairy-tale books, children's book *The Wizard of Oz*, etc, and the skills and tasks involved. It can, thus, be concluded that the examinations have low predictive power to predict performance in the future, in the criterion, thus, with low utility as a university entrance examination.

In short, it seems reasonable to contend that these examinations have sources of invalidity for the interpretation of reading ability, since they may be assessing the writing ability and background knowledge, and also for their use as a university entrance examinations.

6.3 – Consequential Validity: Considerations for the entrance examinations

University entrance examinations in Brazil are high-stakes examinations and have become powerful, with consequences for the society as a whole and for individuals in general. Their power is exercised because they are instruments for the selection of those to be allowed to have access to higher education, thus becoming a screening device, or gatekeeper, to decide the future professionals of the country. Much of what was contributed by Shohamy (2001), discussed in chapter 2, can be applied to these examinations.

In particular, it is possible to claim that tests are administered by powerful institutions which have total control over most decisions such as what and how to test, how to score, and interpret results, use language of science and the language of number, use written communication, rely on documentation, allow for one correct answer determined by the test writer in advance, not open to interpretations. All these characteristics of the institutions turn test takers into powerless individuals. In addition to being powerless, individuals become stigmatized as winners and losers, may develop low self-trust for not having the high-status knowledge, and may have to choose a profession other than the one desired.

In this section, rather than using the validity table for considerations of the arguments as to the consequences of using the entrance examinations, I will provide some reflection based on the analysis of the items.

In the case of UFSC examinations, as to the choice of the texts, it is possible to argue that, although they have topics of general interest, the fact that they refer to the same topic raises an ethical problem, since it might favor those more familiar with

them, thus being biased against those less familiar with them. The consequence of this choice is that, in a selection process such as these entrance examinations, in addition to the reading ability, topical knowledge is also playing a role, being, thus, a source of invalidity based on the validation criteria used in this research.

Some of the items require high contribution from the reader for the inferential process, thus involving reader's background knowledge to a great extent, such as the item with the summaries. However, the fact that they are all items for selected responses is in itself a minimizing factor for the effect of background knowledge for testing, as discussed in chapter 3. Thus, it is possible to argue that the items may not present major problems in terms of the methodological and ethical (fairness) issues as discussed in chapter 3.

In the case of the UNICAMP entrance examination, there are two aspects to be considered: one concerning intended consequences and the other, unintended consequences. Creating a backwash effect is one intended consequence. It is recognized, in the Review Book, that some skills focused on in the examinations are not usually worked upon in the secondary schools, such as contextualizing the origin of texts and recovering the controlling idea of a text (Review Book). A plausible explanation for an entrance examination to assess skills not worked upon previously is that their proponents want to create the backwash effect on the secondary school in such a way that it promotes the kind of reading they advocate.

In fact, Scaramucci (2002) confirms that the realignment of the secondary school curriculum as a backwash effect was a constituent part of the reform proposed for UNICAMP entrance examination. However, in a research carried out with three teachers in different teaching situations, the author found that the backwash effect was not as expected in the three different settings:

The conclusion of this study, aimed at investigating the impact of a reading ability exam on three settings of EFL teaching at the secondary level in Brazil [upper middle-class private school, typical public school, and a high-ranking private extracurricular institution] is that despite the fact that the exam is theoretically oriented towards current views of the reading process and has been in use for over a decade, it failed to change the beliefs of the teachers regarding reading and the teaching of reading, which still seems to be viewed as a passive process of decoding, following the tradition of teaching reading in the mother tongue in first and secondary levels as well as of foreign languages in Brazil (p. 77).

There is an efficacy issue to be discussed based on the conclusion. Backwash effect is not under the control of the test developers, a fact which is not only recognized by the author, when citing conclusions by other researchers in the area, that teachers and learners are affected differently by the backwash effect created by tests, but also a fact that was shown, with her own findings, that the three participants of her research had different influence of the backwash effect.

Scaramucci (2002) provides her interpretation for her findings that the teachers' personal beliefs and educational background may have a stronger effect on the teachers' choices for working with processes of decoding of reading for the classes than the entrance examination itself. This interpretation, of course, deserves investigation and, in my view, within the broader question of why theory and practice are not talking to one another in this matter.

In addition to the intended consequences, there can be arguments for the unintended consequences of discriminating against a group of test takers. Since the least effect of the examination was found to be in the setting of the public school, in the same research, suggesting that public school students have reading lessons in a different perspective from the one advocated by the proponents of UNICAMP entrance examinations, a plausible conclusion is that UNICAMP examinations are discriminating against the students coming from public schools.

Two related questions may be raised: *should tests be used to implement reforms?* and *should tests define or redefine knowledge?* I want to contend that they should not

for the simple reason that this is authoritarian, undemocratic, and unethical, affecting the teachers as well as the secondary students. For that, I resort to Shohamy (2001), who has addressed this issue, and argues against the use of tests as a way of implementing reform in education, comparing the use of tests to improve education to the use of a thermometer to measure the temperature as a way of curing the illness, i.e., very ineffective for solving the problems since it deals with the symptoms only (p. 109).

According to Shohamy (2001), the reason behind the idea of using the test to implement changes as backwash effect is that “teachers and students are not trying hard enough ... [and with] pressure on them through threats or failures, teachers and students would try harder and achieve more” (p. 110), which is, according to the author, compared to informing the teachers and students that they are not doing anything right, and that they will have to change their procedures in accordance to what those in authority believe to be right.

Shohamy (2001) analyzed various situations in which tests were used to bring out changes and concluded that “pressure and sanctions alone are not enough” (p. 111) and that those willing to implement pedagogical changes this way were more interested in “simplistic solutions, where gains can be seen immediately, than in meaningful changes” (p.112). Meaningful changes, according to her, include addressing issues such as class size, reduced workload, and guaranteed training, with workshops, in-service courses, discussion on the nature of methods of teaching, and agreed-upon criteria of quality.

In terms of redefinition of knowledge, Shohamy (2001) claims that high-stakes tests have the power of implementing new knowledge, since they promote certain values and diminish others, upgrade certain abilities and downgrade others, dictating

what schools have to attain, what teachers have to teach, thus becoming the *de facto* curriculum, power which has reached such high levels that its content is believed to be important, and its results indicate status, i.e., where people stand on the valued knowledge. This is, according to her, authoritarian because the knowledge the tests define as valued becomes the institutionalized knowledge.

Assessment instruments designed to assess reading ability should not, in my view, be used to change teachers' beliefs and procedures, and determine school curriculum. Rather, they should assess reading ability the best way theory in the area of testing and reading allows them to. Other forums for the proposal and implementation of the necessary realignment should be created, in which there should be explanation of the underlying reasons for the changes, and supervised and assisted commitment as to the changes to be effected.

Using tests to promote change and to define high-status knowledge and the *de facto* curriculum the way it seems to be the case of UNICAMP examination is, in my view, authoritarian and coercive, since the change and the definitions are dictated from above, without including those affected. In this respect, I agree with Shohamy (2001) when she says that a test, used for promoting changes, narrows “the process of education...making it merely instrumental, and not meaningful” (p. 110).

CHAPTER VII

Conclusions, Final Remarks, Limitations, Implications and Suggestions

In this chapter, I present the main conclusions of the study in section 7.1, final remarks in section 7.2, the limitations of the study in section 7.3, and the major political and pedagogical implications in section 7.4.

7.1 - Conclusions

The objective of the study was to investigate the defensibility of test items used in different testing situations, in terms of the evidential basis for justifications of the interpretation inferences and action inferences, and also in terms of the consequential basis for justifications of interpretation inferences and action inferences.

Based on the validity table used for each of the items analyzed, it is possible to conclude that there are items with only justifications in favor of its use in the test, in particular items assessing lexical cohesion. Also there are items with stronger justifications in favor of their use, such as the items focusing on information recoverable in the text, requiring inference of word meaning and elementary inference of associating proposition in the text with propositions in the items, and items focusing on the identification of the purpose of the writer, requiring recognition of the functional knowledge of text. Also, there are some items focusing on syntactical knowledge for combining sentences. All of them allow for high degree of validity evidence.

There are, however, items with stronger justifications against their use. Four types can be considered extreme cases. The first type refers to items assessing the

writing ability rather than the reading ability, as some of the items used within the UNICAMP entrance examination, in particular, items requiring high level of elaboration for the cognitive operation of explaining. The second type refers to the items assessing the construct of vocabulary knowledge rather than reading ability, as it is possible to see in all the tests examined, either because they are in lists without context, or cannot be inferred within their context. In all the cases, variance can probably be attributed to vocabulary knowledge, which means that the items are assessing vocabulary knowledge rather than reading ability.

Although vocabulary knowledge may be considered a pre-requisite knowledge for reading, as I have shown elsewhere (Tumolo, 1999), the conclusion that a test taker does not know a word given in isolation cannot be generalized to not having the reading ability, specially because language ability has the strategic component of relying on inference making of unknown words.

The third type of items with stronger justifications against them refers to items assessing background knowledge rather than reading ability, as some of the items used within the UNICAMP entrance examination, whose answers rely on scriptal knowledge presupposed by test developers for the test takers. Since variance can probably be attributed to background knowledge, the items are assessing background knowledge rather than reading ability. The fourth type refers to passage-independent items, in the case when the item provides unintended clues, or construct-irrelevant easiness, in the case when the item can be answered based on knowledge of Portuguese, or in the case the test taker knows the answer in advance.

All these items allow for low degree of validity evidence taken as isolated, and would probably converge to low degree of validity of the inferences drawn, based on the test as a whole.

Syntactical knowledge discretely assessed in terms of parsing nominal groups was part of both tests used by the EAP teachers. Although this skill may be subsumed in higher-order skills, hence assessed indirectly, since they are a problem for Brazilian, usually affecting comprehension negatively, they could have been included as the focus of comprehension questions in the other tests analyzed.

Concerning authenticity as investigated through the framework by Bachman (1990) and Bachman and Palmer (1996), it was possible to conclude that some items used within the two entrance examinations focus on characteristics with low degree of authenticity. In particular, it is possible to concluded that the type of texts used have low authenticity in terms of the characteristics required, especially within the characteristics under *language of input*, namely, topical characteristics and language characteristics, in particular organizational characteristics (textual) and pragmatic characteristics (functional and sociolinguistic). This low degree of correspondence allows little room for generalization from performance in the test to performance in the criterion.

Considering consequential validity, or impact as called in Bachman and Palmer (1996), it was possible to conclude that some items used within the UNICAMP entrance examination have what can be argued to be unintended consequence of discriminating against groups based on factors irrelevant to the construct, based on specific knowledge the developers consider relevant. This is a fact that outweighs what can be argued to be the intended consequence of university entrance examinations, i.e., to discriminate between those with the ability and those without it for admission to the criterion.

Since background knowledge plays its most important role in the inferential processes, using items focusing on these knowledge-based processes will always raise

the validity question of what is the underlying knowledge that the item is assessing. Assessing scriptal knowledge or specific content knowledge by the use of pragmatic inferencing, as discussed in chapter 3, raises the ethical question of who has the 'correct understanding' to be used as reference for interpretation inferences (interpretation of reading ability) and for action inferences (allow or deny access to the criterion).

It may be a safer ground, methodologically and ethically speaking, to use items assessing organizational knowledge in terms of grammatical knowledge (vocabulary, syntax) and textual knowledge (cohesion, rhetorical organization), and pragmatic knowledge in terms of the functional knowledge (ideational functions, in particular), and definitely leave out sociolinguistic knowledge, which requires knowledge with cultural references. All in all, they are reading tests, not knowledge tests.

7.2 - Final Remarks

This research has contributed the most current definitions of validity, adopted, in particular, by American scholars in the area of testing, as well as the most recent perspectives on how to assess the degree of the validity of the interpretations to be drawn and actions to be made based on the performance elicited by test items. It has contributed to the learning of what methods have been used to assess reading comprehension in some testing situations, and to the understanding of the possible rationale behind the choices.

My main objective was to provide a starting point for the discussion of how to assess reading ability, as pointed out in the chapter on the method. Although I have provided definite answers to the question on the defensibility of each item, I have also

provided the balance of arguments. There is seldom a test item which is absolute in its defensibility, in particular because analysis of defensibility is based on constructs, and constructs are subject to changes, resulting in changes in the arguments.

During the analysis of the items, I noticed that the process of validation involving arguments for a validity conclusion is not a simple task. First, the consideration of different arguments would possibly lead to a different validation conclusion, sometimes opposite to the previous one. In this respect, I share this feeling of venture with Kane (1992) when he claims that validation conclusion is dynamic.

Second, providing the appropriate type and number of justifications is rather challenging in all the phases, starting with the hypothesis for argument in favor, which is the result of what you imagine was the rationale for the development of the item and what can be deduced from the construct of reading as used in this research, and then carrying on to the next steps of finding arguments against and the possible refutations of the arguments against, both steps to investigate the force of the arguments. In this respect, I share this feeling of challenge with Chapelle (1999), when she recognizes the difficulties involved in finding the types and numbers of justifications and in integrating them for a validation conclusion. This challenge must, in my view, be seen as part of the process, particularly in these more interpretive methods of investigation.

My final remarks concerning the conclusions of the research are based on two of the principles of language testing proposed by Bachman and Palmer (1996), which they call *our philosophy of language testing*. The first principle is: “design your tests so as to encourage and enable test takers to perform at their highest level of ability” (p. 18). Developing items with unclear instructions, with trick questions, with presupposed knowledge, with construct- and criterion- irrelevance, with low authenticity, as found in this study, goes against this principle.

The second principle is: “build considerations of fairness into test design” (Bachman & Palmer, 1996, p. 18). Fairness is related to the ethical issue of bias, and it addresses the question of what knowledge is presupposed in the items, based on whom, favoring whom, against whom. Some items used in the examinations go against this principle.

Fairness also addresses the issue of the power of tests, in particular in the case of high-stakes selection tests. Based on the discussion by Shohamy (2001), presented in chapter 2, it is possible to say that university entrance examinations have some built-in features of power. They: regulate behavior, deciding who passes and who fails; identify test takers as scores, classifying them as success or failure; are used for judgment and sanctions and for declaring where authority lies; and are kept secret. They, thus, cause a lot of tension, fear, and anxiety among test takers.

And since affective schemata, as described by Bachman and Palmer (1996), affects performance negatively, in particular the strategic component (Bachman & Palmer, 1996), it is high time we do our best to change that. I hope this research can contribute to that pursuit.

7.3 - Limitations of the Research

I recognize that the method of expert judgment providing the self-report protocol in probing the items to raise hypothesis of skills or levels assessed by some items is not simple, since the performance elicited by the items most likely involve a combination of skills. My expert judgment is only a starting point for further research.

Also, I acknowledge that the process of validation is ongoing, and validation as argument-based is also dynamic, in that it may change with the introduction of any new argument, with the change of any aspect considered in the construct. The ‘yes’ or ‘no’ as referring to the defensibility of each item are only relative to the arguments, being subject to change in case the arguments change. This means that each item will be defensible only to the extent of the plausibility of the arguments presented, and can only be considered defensible or non-defensible together with the arguments.

Moreover, as already stressed, validity is a matter of degree, not an all-or-nothing thing. The validity of the interpretation and action based on the items will be relative to the strength of the argumentation, being more or less valid, considering all the arguments presented.

It is important to mention that the present discussion is related to the situation of reading expository texts with the purpose of studying, which is the target language use situation for all the three testing situations (the criterion). Different analysis is required for different reading purposes and different text types.

The present discussion is related to tests items for use with Brazilians as test takers, with Portuguese as their native language, which shares the same alphabet and has many cognate words. Different analysis is required for test takers whose native languages do not share the same alphabet or cognate words.

It is, also, related to tests used with groups heterogeneous in terms of background knowledge. Different analysis is required for test takers whose background knowledge or topical knowledge is similar.

In addition, it is related to tests and test items within traditional testing. Different analysis is required for alternative assessment procedures such as portfolios, self-assessment, projects, observations, or critical testing as advocated by Shohamy (2001). Also, it is related to the assessment of reading as a discrete skill. Different analysis is required in case of more integrated skills within a communicative approach.

Equally important is to stress that the main conclusions refer to language testing, more specifically reading assessment, since the data was collected in language testing situations. Despite this limitation, the whole discussion of validity, and the process of collecting evidence and providing argument, in the attempt to support the inferences and actions, can be applied to any testing situation.

7.4 - Political and Pedagogical Implications

The political implication is the claim for the development of any testing procedure involving, in terms of evidential basis, the assessment and clarification of the construct to be assessed and the methods used for its assessment, and, in terms of consequential basis, the accountability of the consequences, both intended and unintended, of the testing procedures used, particularly concerning test bias. The pursuit to avoid bias necessarily involves discussions on what to test, how to test, and how to score. But most importantly, it involves discussions on, and accountability of, the underlying reasons for all the decisions taken, all grounded on, and, at the same time revealing, the model of society each test user is seeking for. The pedagogical

implication is that it provides invaluable information as guidelines for any teacher to develop their tests in any area within educational settings. There is also a direct pedagogical implication, since it contributes for the EAP teachers to develop their testing procedures. Ultimately, the contribution of this study is in hope of the development of testing procedures with features of validity, which may change their image among test takers, changing its ugly history, as expressed by those kids in the film 'Harry Potter', and changing, also, its role in our society from instruments of power and gatekeeping to instruments with features for the evaluation of the whole process of education.

7.5 - Suggestions for Further Research

The analysis in this research is based only on the judgment of the researcher as an expert judge. Further research involving data collected based on the responses given by participants, in particular students, will add to the arguments presented in the research. Using, with the data collected, Classical Item Analysis with the two indexes of facility value and discrimination index, Item Response Theory with the one-parameter, two-parameter, and three-parameter models, and Descriptive statistics, as well as more qualitative methods of introspection and retrospection, all mentioned by Alderson (2000), will provide more evidence for the validation study carried out here.

REFERENCES

- ACTFL Proficiency Guidelines (1986). American Council on the Teaching of Foreign Languages. New York: Hastings-on-Hudson.
- Aebersold, J. & Field, M. (1997). *From reader to reading teacher*. New York: CUP.
- Alderson, J. C. (1984). Reading: A reading problem or a language problem? In C. Alderson & H. Urquhart (Eds.), *Reading in a foreign language*. Longman.
- Alderson, J. C. (1990a). Testing reading comprehension skills (part one). *Reading in a foreign language* 6 (2), 425-438.
- Alderson, J. C. (1990b). Testing reading comprehension skills (part two). *Reading in a foreign language* 7 (1), 465-503.
- Alderson, J. C. (1996). The testing of reading. In C. Nuttall (1996), *Teaching reading skills in a foreign language* (new edition). Oxford, UK: Macmillan Education.
- Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.
- Alderson, J. C. (1999). Reading construct and reading assessment. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency*. Cambridge, UK: Cambridge University Press.
- Alderson, J. C. & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language* 5 (2), 253-270.
- Alderson, J. C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, UK: Cambridge University Press.
- Allison, D. (1999). *Language testing and evaluation*. Singapore: Singapore University Press.
- Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology* 37, 1-15.
- Ashcraft, M. (1994). *Human memory and cognition*. HarperCollins College Publishers.

- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.
- Bachman, L. & Palmer, A. (1982). The construct validation of some components of some components of communicative proficiency. *TESOL Quarterly* 16 (4), 449-465.
- Bachman, L., Davidson, F., Ryan, K. & Choi, I. (1995). *An investigation into the comparability of two tests of English as a foreign language*. Cambridge, UK: Cambridge University Press.
- Bachman, L., Davidson, F. & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing*, 13 (2).
- Bachman, L. & Palmer, A. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.
- Bachman, L. & Cohen, A. (1998). *Interfaces between second language acquisition and language testing research*. Cambridge, UK: Cambridge University Press.
- Ballstaedt, S. & Mandl, H. (1984). Elaborations: Assessment and analysis. In H. Mandl, N.L. Stein & T. Trabasso (Eds.), *Learning and comprehension of text*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Bernhardt, E. (1986). Proficiency Texts or Proficient Readers? *ADFL Bulletin* 18 (1), 25-28.
- Bernhardt, E. (1991). *Reading development in a second language*. Norwood, NJ: Ablex Publishing Corporation.
- Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research*. Cambridge, UK: Cambridge University Press.
- Brindley, G. (2001). Assessment. In R. Carter & D. Nunan (Eds.), *The Cambridge guide to teaching of English to speakers of other languages*. Cambridge, UK: Cambridge University Press.
- Brown, D. (1987). *Principles of language learning and teaching*. Englewood Cliffs, NJ: Prentice Hall Regents.

- Brown, D. (2000). *Assessing languages for specific purposes*. Cambridge, UK: Cambridge University Press.
- Buffa, L. & Pearson, L. (1998). *Cracking the TOEFL*. Princeton Review Publishing, L.L.C.
- Byrnes, H. (1986). Proficiency: Concepts and developments. *ADFL Bulletin* 18 (1), 9-10.
- Cambridge University Press (2004). *Cambridge IELTS 3: Examination papers from university of Cambridge ESOL examinations* (4th ed.). Cambridge: Cambridge University Press.
- Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1 (1), 1-47.
- Carrell, P. (1991). Second language reading: Reading ability or language proficiency. In *Applied Linguistics*, 12 (2). Oxford, UK: Oxford University Press.
- Carver, R. P. (1981). *Reading comprehension and reading theory*. Charles C. Thomas Publisher.
- Chapelle, C. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research*, 10 (2), 157-187.
- Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research*. Cambridge, UK: Cambridge University Press.
- Chapelle, C. (1999). Validity in language assessment. *Annual Review of Applied linguistics* 19, 254-272. Cambridge University Press.
- Chastain, K. (1989). The ACTFL Guidelines: A selected sample of opinions. *ADFL Bulletin* 20 (2), 47-51.
- Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge, UK: Cambridge University Press.
- Cronback, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Brown (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum.

- Cumming, A. (1996). Introduction: The concept of validation in language testing. In A. Cumming & R. Berwick (Eds.), *Validation in language testing*. UK: Cromwell Press.
- Davidson, F. & Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven: Yale University Press.
- Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Dictionary of language testing*. Cambridge, UK: Cambridge University Press.
- Davies, F. (1995). *Introducing reading*. London: Penguin Books.
- Douglas, D. (1995). Developments in language testing. *Annual review of applied linguistics* 15, 167-187.
- ETS (1995). *TOEFL Sample test*. Princeton, NJ: Educational Testing Service.
- Faerch, C. & Kasper, G. (1983). *Strategies in interlanguage communication*. London: Longman.
- Foucault, M. (1979). *Discipline and punish: The birth of prison*. New York: Vintage Books
- Gagné, E., Yekovich, C. & Yekovich, F. (1993). *The cognitive psychology of school learning*. New York: Harper Collins College Publishers.
- Gear, J. & Gear, R. (2002). *Cambridge preparation for the TOEFL test*. Cambridge, UK: Cambridge University Press.
- Genesee, F. (2001). Evaluation. In R. Carter & D. Nunan (Eds.), *The Cambridge Guide to Teaching of English to Speakers of other Languages*. Cambridge, UK: Cambridge University Press.
- Grabe, W. (1999). Developments in reading research and their implications for computer-adaptive reading assessment. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency*. Cambridge, UK: Cambridge University Press.
- Grabe, W. (2000). Reading research and its implications for reading assessment. In A. J. Kunnan (Ed.) *Fairness and validation in language assessment*. Cambridge, UK: Cambridge University Press.

- Grellet, F. (1981). *Developing reading skills: A practical guide to reading comprehension exercises*. Cambridge, UK: Cambridge University Press.
- Hale, G. (1988). The interaction of student major-field group and text content in TOEFL reading comprehension. *Research Reports*. Princeton, NJ: Educational Testing Service.
- Heaton, J.B. (1988). *Writing English language tests*. (New ed.). New York: Longman Inc.
- Howe, M. (1996). Concepts of ability. In I. Dennis & P. Tatsfield (Eds.), *Human abilities: their nature and measurement*. New Jersey: LEA.
- Hughes, A. (1989). *Testing for language teachers*. Cambridge, UK: Cambridge University Press.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed). Cambridge, UK: Cambridge University Press.
- Hummel, R. (1985). Evaluating Proficiency in Comprehension Skills: How can we observe what can't be observed? *ADFL Bulletin 16* (2) January. 13-16.
- Just, M. & Carpenter, P. (1984). Reading skills and skilled reading in the comprehension of text. In H. Mandl, N.L. Stein & T. Trabasso (Eds.), *Learning and comprehension of text*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112 (3) 527-535.
- Kasai, M. (1997). *Application of the rule space model to the reading comprehension section of the test of English as a foreign language (TOEFL)*. Unpublished PhD dissertation, University of Illinois at Urbana-Champaign.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: CUP.
- Kintsch, W. & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85.
- Kunnan, A. J. (1998). Approaches to validation in language assessment. In A. J. Kunnan (Ed.), *Validation in language assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

- Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. McGraw-Hill Book Company.
- Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York: Farrar, Straus and Giroux.
- Lorch Jr., R., Klusewitz, M. & Lorch, E. (1995). Distinctions among reading situations. In R. Lorch & E. O'Brien (Eds.), *Sources of coherence in reading*. New Jersey: LEA.
- Matthews, M. (1990). Skill taxonomies and problems for the testing of reading. *Reading in a foreign language*, 7 (1), 511-517.
- McNamara, T. (2000). *Language testing*. Oxford, UK: Oxford University Press.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement*. Washington, DC: The American Council on Education and the National Council on Measurement in Education.
- Norton, B. & Stein, P. (1998). Why the 'Monkeys Passage' bombed: Tests, genres, and teaching. In A. J. Kunnan (Ed.), *Validation in language assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.
- Nuttall, C. (1996). *Teaching reading skills in a foreign language* (new edition). Oxford, UK: Macmillan Education.
- Ommagio Hadley, A. (1993). *Teaching language in context*. Boston, Massachusetts: Heinle & Heinle Publishers.
- Pearson, P. D. & Johnson, D. D. (1978). *Teaching reading comprehension*. New York: Holt, Rinehart and Winston.
- Pierce, B. N. (1992). Demystifying the TOEFL reading test. *TESOL Quarterly*, 26 (4), 665-689.
- Popham, W. J. (1981). *Modern educational measurement*. Englewood Cliffs, NJ: Prentice Hall.
- Scaramucci, M. (2002). Entrance examinations and TEFL in Brazil: A case study. *Revista Brasileira de Lingüística Aplicada*, 2 (1). Belo Horizonte: Faculdade de Letras da UFMG.

- Shepard, L. A. (1993). *Evaluating test validity*. *Review of Research in Education*, 19, 405-450.
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Pearson Education Limited.
- Spiro, R.J., Bruce, B.C. & Brewer, W.F. (1980). *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Tapiero, I. & Otero, J. (1999). Distinguishing between textbase and situation model in the processing of inconsistent information: Elaboration and tagging. In H. Van Oostendorp & S. Goldman (Eds.), *The construction of mental representation during reading*. New Jersey: LEA.
- Tomitch, L. (1991). Schema activation and text comprehension. *Fragmentos*, 3 (2), 9-43.
- Tomitch, L. (2000). Designing reading tasks to foster critical thinking. *Ilha do Desterro* 38, 83-90. Florianópolis, SC: Editora da UFSC.
- Tumolo, C. (1999). *Vocabulary Instruction: The text as a source in the classroom*. Unpublished Thesis, Universidade Federal de Santa Catarina.
- Urquhart, S. & Weir, C. (1998). *Reading in a second language: Process, product and practice*. Longman.
- Weir, C. (1993). *Understanding & developing language tests*. Prentice Hall International
- Weir, C., Hughes, A. & Porter, D. (1990). Reading skills: Hierarchy, implicational relationships and identifiability. *Reading in a foreign language*, 7, 1.
- Weir, C., Huizhong, Y. & Yan, J. (2000). *An empirical investigation of the componentiality of L2 reading in English for academic purposes*. Cambridge, UK: Cambridge University Press.
- Weir, C. & Milanovic, M. (2003). *Continuity and innovation: Revising the Cambridge Proficiency Examination 1913-2002*. Cambridge, UK: Cambridge University Press.

- Wielewicki, G. (1997). *ESP reading proficiency tests: What if ends do not meet?* Paper presented at XIV ENPULI.
- Willingham, W. (1999). A systematic view of test fairness. In S. Messick (1999), *Assessment in higher education: Issues of access, quality, student development, and public policy*. Lawrence Erlbaum Associates, Publishers.
- Zwaan, R. & Brown, C. (1996). The influence of language proficiency and comprehension skills on situation-model construction. *Discourse Processes*, 21.

APPENDIX 1
IELTS Examination – test 1

Reading passage 1

Questions 1-4

Reading Passage 1 has six paragraphs **A-F**.

Choose the most suitable headings for paragraphs **B-E** from the list of headings below.

Write the appropriate numbers **i-ix** in boxes 1-4 on your answer sheet.

List of Headings

- i** How the reaction principle works
- ii** The impact of the reaction principle
- iii** Writers' theories of the reaction principle
- iv** Undeveloped for centuries
- v** The first rockets
- vi** The first use of steam
- vii** Rockets for military use
- viii** Developments of fire
- ix** What's next?

<i>Example</i> Paragraph A	<i>Answer</i> ii
-------------------------------	----------------------------

- 1** Paragraph B
- 2** Paragraph C
- 3** Paragraph D
- 4** Paragraph E

<i>Example</i> Paragraph F	<i>Answer</i> ix
-------------------------------	----------------------------

THE ROCKET – FROM EAST TO WEST

- A** The concept of the rocket, or rather the mechanism behind the idea of propelling an object into the air, has been around for well over two thousand years. However, it wasn't until the discovery of the reaction principle, which was the key to space travel and so represents one the great milestones in the history of

scientific thought, the rocket technology was able to develop. Not only did it solve a problem that had intrigued man for ages, but more importantly, it literally opened the door to exploration of the universe.

- B** An intellectual breakthrough, brilliant though it may be, does not automatically ensure that the transition is made from theory to practice. Despite the fact that rockets had been used sporadically for several hundred years, they remained a relatively minor artefact of civilisation until the twentieth century. Prodigious efforts, accelerated during two world wars, were required before the technology of primitive rocketry could be translated into the reality of sophisticated astronauts. It is strange that the rocket was generally ignored by writers of fiction to transport their heroes to mysterious realms beyond the Earth, even though it had been commonly used in fireworks displays in China since the thirteenth century. The reason is that nobody associated the reaction principle with the idea of travelling through space to a neighbouring world.
- C** A simple analogy can help us to understand how a rocket operates. It is much like a machine gun mounted on the rear of a boat. In reaction to the backward discharge of bullets, the gun, and hence the boat, move forwards. A rocket motor's 'bullets' are minute, high-speed particles produced by burning propellants in a suitable chamber. The reaction to the ejection of these small particles causes the rocket to move forwards. There is evidence that the reaction principle was applied practically well before the rocket was invented. In his *Noctes Atticae* or *Greek Nights*, Aulus Gellius describes 'the pigeon of Archytas', an invention dating back to about 360 BC. Cylindrical in shape, made of wood, and hanging from string, it was moved to and fro by steam blowing out from small exhausted ports at either end. The reaction to the discharging steam provided the bird with motive power.
- D** The invention of rockets is linked inextricably with the invention of 'black powder'. Most historians of technology credit the Chinese with its discovery. They base their belief on studies of Chinese writings or on notebooks of early Europeans who settled in or made long visits to China to study its history and civilisation. It is probably that, some time in the tenth century, black powder was first compounded from its basic ingredients of saltpeter, charcoal and sulphur. But this does not mean that it was immediately used to propel rockets. By the thirteenth century, powder-propelled fire arrows had become rather common. The Chinese relied on this type of technological development to produce incendiary projectiles of many sorts, explosive grenades and possibly cannons to repel their enemies. On such weapon was the 'basket of fire' or, as directly translated from Chinese, the 'arrows like flying leopards'. The 0.7 metre-long arrows, each with a long tube of gunpowder attached near the point of each arrow, could be fired from a long, octagonal-shaped basket at the same time and have a range of 400 paces. Another weapon was the 'arrow as a flying sabre', which could be fired from crossbows. The rocket, placed in a similar position to other rocket-propelled arrows, was designed to increase the range. A small iron weight was attached to the 1.5m bamboo shaft, just below the feathers, to increase the arrow's stability by moving the center of gravity to a position below the rocket. At a similar time, the Arabs had developed the 'egg

which moves and burns'. This 'egg' was apparently full of gunpowder and stabilised by a 1.5m tail. It was fired using two rockets attached to either side of this tail.

- E** It was not until the eighteenth century that Europe became seriously interested in the possibilities of using the rockets itself as a weapon of war and not just to propel other weapons. Prior to this, rockets were used only in pyrotechnic displays. The incentive for the more aggressive use of rockets came not from within the European continent but from far-away India, whose leaders had built up a corps of racketeers and use rockets successfully against the British were described by a British Captain serving in India as 'an iron envelope about 200 millimetres long and 40 millimetres in diameter with sharp points at the top and 3m-long bamboo guiding stick'. In the early nineteenth century the British began to experiment with incendiary barrage rockets. The British rocket differed from the Indian version in that it was completely encased in a stout, iron cylinder, terminating in a conical head, measuring one metre in diameter and having a stick almost five metres long and constructed in such a way that it could be firmly attached to the body of the rocket. The Americans developed a rocket, complete with its own launcher, to use against the Mexicans in the mid-nineteenth century. A long cylindrical tube was propped up by two sticks and fastened to the top of the launcher, thereby allowing the rockets to be inserted and lit from the other end. However, the results were sometimes not that impressive as the behaviour of the rockets in flight was less than predictable.
- F** Since then, there have been huge developments in rocket technology, often with devastating results in the forum of war. Nevertheless, the modern day space programs owe their success to the humble beginnings of those in previous centuries who developed the foundations of the reaction principle. Who knows what it will be like in the future?

Questions 5 and 6

Choose the appropriate letters A-D and write them in boxes 5 and 6 on your answer sheet.

- 5** The greatest outcome of the discovery of the reaction principle was that
- A** rockets could be propelled into the air.
 - B** space travel became a reality.
 - C** a major problem had been solved.
 - D** bigger rockets were able to be built.
- 6** According to the text, the greatest progress in rocket technology was made
- A** from the tenth to the nineteenth centuries.
 - B** from the seventeenth to the nineteenth centuries.
 - C** from the early nineteenth to the late nineteenth century.
 - D** from the late nineteenth century to the present day.

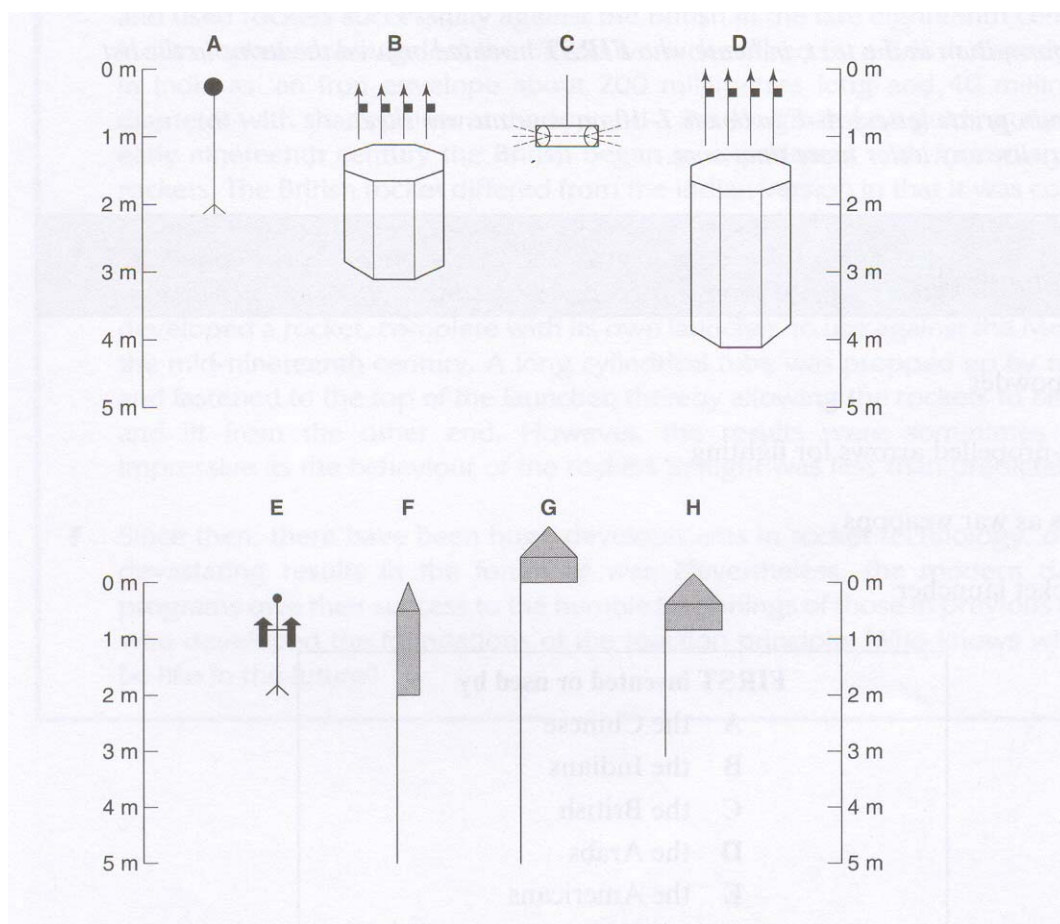
Questions 11-14

Look at the drawings of different projectiles below, **A-H**, and the names of types of projectiles given in the passage, **Questions 11-14**. Match each name with one drawing.

Write the appropriate letters **A-H** in boxes 11-14 on your sheet.

<i>Example</i>	<i>Answer</i>
The Greek 'pigeon of Archytas'	C

- 11** The Chinese 'basket of fire'
12 The Arab 'egg which moves and burns'
13 The Indian rocket
14 The British barrage rocket



Reading passage 2

You should spend about 20 minutes on **Questions 15-28** which are based on Reading Passage 2 below.

The Risks of Cigarette Smoke

Discovered in the early 1800s and named nicotianine, the oily essence now called nicotine is the main active ingredient of tobacco. Nicotine, however, is only a small component of cigarette smoke, including 43 cancer-causing substances. In recent times, scientific research has been providing evidence that years of cigarette smoking vastly increases the risk of developing fatal medical conditions.

In addition to being responsible for more than 85 per cent of lung cancers, smoking is associated with cancers of, amongst others, the mouth, stomach and kidneys, and is thought to cause about 14 per cent of leukemia and cervical cancers. In 1990, smoking caused more than 84,000 deaths, mainly resulting from such problems as pneumonia, bronchitis and influenza. Smoking, it is believed, is responsible for 30 per cent of all deaths from cancer and clearly represents the most important preventable cause of cancer in countries like the United States today.

Passive smoking, the breathing in of the side-stream smoke from the burning of tobacco between puffs or of the smoke exhaled by a smoker, also causes a serious health risk. A report published in 1992 by the US Environmental Protection Agency (EPA) emphasized the health dangers, especially from side-stream smoke. This type of smoke contains more, smaller particles and is therefore more likely to be deposited deep in the lungs. On the basis of this report, the EPA has classified environmental tobacco smoke in the highest risk category for causing cancer.

As an illustration of the health risks, in the case of a married couple where one partner is a smoker and one a non-smoker, the latter is believed to have a 30 per cent higher risk of death from heart disease because of passive smoking. The risk of lung cancer also increases over the years of exposure and the figure jumps to 80 per cent if the spouse has been smoking four packs a day for 20 years. It has been calculated that 17 per cent of cases of lung cancer can be attributed to high levels of exposure to second-hand tobacco smoke during childhood and adolescence.

A more recent study by researchers at the University of California at San Francisco (UCSF) has shown that the second-hand cigarettes smoke does more harm to non-smokers than to smokers. Leaving aside the philosophical question of whether anyone should have to breath someone else's cigarette smoke, the report suggests that the smoke experienced by many people's heart and lungs.

The report, published in the Journal of American Medical Association (AMA), was based on the researchers' own earlier research but also include a review of the studies over the past few years. The American medical Association represents about half of all US doctors and is a strong opponent of smoking. The study suggests that people who smoke cigarettes are continually damaging their cardiovascular system, which adapts in order to compensate for the effects of smoking. It further states that people who do not smoke do not have the benefit of their system adapting to the smoking inhalation. Consequently, the effects of passive smoking are far greater on non-smokers than on smokers.

The report emphasizes that cancer is not caused by a single element in cigarette smoke; harmful effects to health are caused by many components. Carbon monoxide, for example, combines with oxygen in red blood cells and interferes with the blood's ability to deliver life-giving oxygen to the heart. Nicotine and other toxins in cigarette smoke activate small blood cells called platelets, which increases the likelihood of blood clots, thereby affecting blood circulation throughout the body.

The researchers criticize the practice of some scientific consultants who work with tobacco industry for assuming that cigarette smoke has the same impact in smokers as it does on non-smokers. They argue that those scientists are underestimating the damage done by passive smoking and, in support of their recent findings, cite some previous research which points to passive smoking as the cause for the between 30,000 and 60,000 deaths from heart attacks each year in the United States. This means that passive smoking is the third most preventable cause of death after active smoking and alcohol-related diseases.

The study argues that this type of action needed against passive smoking should be similar to that being taken against illegal drugs and AIDS (SIDA). The UCSF researchers maintain that the simplest and most cost-effective action is to establish smoke-free work places, schools and public places.

Questions 15-17

Choose the appropriate letters A-D and write them in boxes 15-17 on your answer sheet.

- 15** According to information in the text, leukemia and pneumonia
- A** are responsible for 84,000 deaths each year.
 - B** are strongly linked to cigarette smoking.
 - C** are strongly linked to lung cancer.
 - D** result in 30 per cent of deaths per year.
- 16** According to information in the text, intake of carbon monoxide
- A** inhibits the flow of oxygen to the heart.
 - B** increases absorption of other smoke particles.
 - C** inhibits are blood cell formation.
 - D** promotes nicotine absorption.
- 17** According to information in the text, intake of nicotine encourages
- A** blood circulation through the body.
 - B** activity of other toxins in the blood.
 - C** formation of blood clots.
 - D** an increase of platelets in the blood.

Question 18-21

Do the following statements reflect the claims of the writer in Reading Passage 2?

In boxes 18-21 on your answer sheet write

YES *if the statement reflects the claims of the writer*
NO *if the statement contradicts the claims of the writer*
NOT GIVEN *if it is impossible to say what the writer thinks about this*

- 18** Thirty per cent of deaths in the United States are caused by smoking-related diseases.
- 19** If one partner in a marriage smokes, the other is likely to take up smoking.
- 20** Teenagers whose parents smoke are at risk of getting lung cancer at some time during their lives.
- 21** Opponents of smoking financed the UCSF study.

Questions 22-24

*Choose **ONE** phrase from the list of phrases **A-J** below to complete each of the following sentences (Questions 22-24).*

Write the appropriate letters in boxes 22-24 on your answer sheet.

- 22** Passive smoking...
- 23** Compared with a non-smoker, a smoker...
- 24** The American Medical Association...

- | |
|---|
| <p>A includes reviews of studies in its reports.</p> <p>B argues for stronger action against smoking in public places.</p> <p>C is one of the two most preventable causes of death.</p> <p>D is more likely to be at risk from passive smoking diseases.</p> <p>E is more harmful to non-smokers than to smokers.</p> <p>F is less likely to be at risk of contracting lung cancer.</p> <p>G is more likely to be at risk of contracting various cancers.</p> <p>H opposes smoking and publishes research on the subject.</p> <p>I is just as harmful to smokers as it is to non-smokers.</p> <p>J reduces the quantity of blood flowing around the body.</p> |
|---|

Questions 25-28

Classify the following statements as being

- A** a finding of the UCSF study
- B** an opinion of the UCSF study
- C** a finding of the EPA report
- D** an assumption of the consultants to the tobacco industry

Write the appropriate letters **A-D** in boxes 25-28 on your answer sheet.

NB You may use any letter more than once.

- 25** Smokers' cardiovascular systems adapt to the intake of environment smoke.
- 26** There is a philosophical question as to whether people should have to inhale others' smoke.
- 27** Smoke-free public places offer the best solution.
- 28** The intake of side-stream smoke is more harmful than smoke exhaled by a smoker.

Reading passage 3

Questions 29-33

Reading Passage 3 has seven paragraphs A-G.

Choose the most suitable headings for paragraphs C-G from the list of headings below.

Write the appropriate numbers I-x in boxes 29-33 on your answer sheet.

List of Headings

- i** The Crick and Watson approach to research
- ii** Antidotes to bacterial infection
- iii** The testing of hypotheses
- iv** Explaining the inductive method

- v Anticipating results before data is collected
- vi How research is done and how it is reported
- vii The role of hypotheses in scientific research
- viii Deducing the consequences of hypotheses
- ix Karl Popper's claim that scientific method is hypothetico-deductive
- x The unbiased researcher

<i>Example</i>	<i>Answer</i>
Paragraph A	ix

- 29 Paragraph C
- 30 Paragraph D
- 31 Paragraph E
- 32 Paragraph F
- 33 Paragraph G

THE SCIENTIFIC METHOD

- A** 'Hypotheses,' Said Medawar in 1964, 'are imaginative and inspirational in character'; they are 'adventures of the mind'. He was arguing in favour of the position taken by Karl Popper in *The Logic of Scientific Discovery* (1972, 3rd edition) that the nature of scientific method is hypothetico-deductive and not, as is generally believed, inductive.
- B** It is essential that you, as an intending researcher, understand the difference between these two interpretations of the research process so that you do not become discouraged or begin to suffer from a feeling of 'cheating' or not going about it the right way.
- C** The myth of scientific method is that it is inductive: that the formulation of scientific theory starts with the basic, raw evidence of the senses – simple, unbiased, unprejudiced observation. Out of these sensory data – commonly referred to as 'facts' – generalizations will form. The myth is that from a disorderly array of factual information an orderly, relevant theory will somehow emerge. However, the starting point of induction is an impossible one.
- D** there is no such thing as an unbiased observation. Every act of observation we make is a function of what we have seen or otherwise experienced in the past. All scientific work of an experimental or exploratory nature starts with some expectation about the outcome. This expectation is a hypothesis. Hypotheses provide the initiative and incentive for the inquiry and influence the method. It is in the light of an expectation that some observations are held to be relevant and some irrelevant, that one methodology is chosen and others discarded, that some experiments are conducted and others are not. Where is your naïve, pure and objective researcher now?

- E** Hypotheses arise by guesswork, or inspiration, but having been formulated they can and must be tested rigorously, using the appropriate methodology. If the predictions you make as a result of deducting certain consequences from your hypothesis are not shown to be correct then you discard or modify your hypothesis. If the predictions turn out to be correct then your hypothesis has been supported and may be retained until such time as some further test shows it not to be correct. Once you have arrived at your hypothesis, which is a product of your imagination, you then proceed to a strictly logical and rigorous process, based upon deductive argument – hence the term ‘hypothetico-deductive’.
- F** So don’t worry if you have some idea of what your results will tell you before you even begin to collect data; there are no scientists in existence who really wait until they have all the evidence in front of them before they try to work out what it might possibly mean. The closest we ever get to this situation is when something happens by accident; but even then the researcher has to formulate a hypothesis to be tested before being sure that, for example, a mould might prove to be a successful antidote to bacterial infection.
- G** The myth of scientific method is not only that it is intuitive (which we have seen is incorrect) but also that the hypothetico-deductive method proceeds in a step-by-step, inevitable fashion. The hypothetico-deductive method describes the logical approach to much research work, but it does not describe the *psychological* behaviour that brings it about. This is much more holistic – involving guesses, reworkings, corrections, blind alleys and above all inspiration, in the deductive as well as the hypothetic component – than is immediately apparent from reading the final thesis or published papers. These have been, quite properly, organized into a more serial, logical order so that the worth of the *output* may be evaluated independently of the behavioural processes by which it was obtained. It is the difference, for example between the academic papers with which Crick and Watson demonstrated the structure of the DNA molecule and the fascinating book *The Double Helix* in which Watson (1968) described how they did it. From this point of view, ‘scientific method’ may more usefully be thought of as a way of *writing up* research rather than as a way of carrying it out.

Questions 34 and 35

*In which **TWO** paragraphs in Reading Passage 3 does the writer give advice **directly** to the reader?*

*Write the **TWO** appropriate letters (A-G) in boxes 34 and 35 on your answer sheet.*

Questions 36-39

Do the following statements reflect the opinions of the writer in Reading Passage 3?

In boxes 36-39 on your answer sheet write

YES if the statement reflects the opinion of the writer
NO if the statement contradicts the opinion of the writer
NOT GIVEN if it is impossible to say what the writer thinks about this

- 36** Popper says that the scientific method is hypothetico-deductive.
- 37** If a prediction based on a hypothesis is fulfilled, then the hypothesis is confirmed as true.
- 38** Many people carry out research in a mistake way.
- 39** The 'scientific method' is more a way of describing research than a way of doing it.

Question 40

*Choose the appropriate letter **A-D** and write it in box 40 on your answer sheet.*

Which of the following statements best describes the writer's main purpose in Reading Passage 3?

- A** to advise Ph.D students not to cheat while carrying out research
B to encourage Ph.D students to work by guesswork and inspiration
C to explain to Ph.D students the logic which the scientific research paper follows
D to help Ph.D students by explaining different conceptions of the research process.

APPENDIX 2

IELTS Examination – test 2

Reading passage 1

You should spend about 20 minutes on Questions 1-13 which are based on Reading Passage 1 below.

A Remarkable Beetle

Some of the most remarkable beetles are the dung beetles, which spend almost their whole lives eating and breeding in dung¹.

More than 4,000 species of these remarkable creatures have evolved and adapted to the world's different climates and the dung of its many animals. Australia's native dung beetles are scrub and woodland dwellers, specializing in coarse marsupial droppings and avoiding the soft cattle dung in which bush flies and buffalo flies breed.

In the early 1960s George Bornemissza, then a scientist at the Australian Government's premier research organization, the Commonwealth Scientific and Industrial Research Organisation (CSIRO), suggested that dung beetles should be introduced to Australia to control dung-breeding flies. Between 1968 and 1982, the CSIRO imported insects from about 50 different species of dung beetle, from Asia, Europe and Africa, aiming to match them to different climatic zones in Australia. Of the 26 species that are known to have become successfully integrated into the local environment, only one, an African species released in northern Australia, has reached its natural boundary.

Introducing dung beetles into a pasture is a simple process: approximately 1,500 beetles are released, a handful at a time, into fresh cow pats² in the cow pasture. The beetles immediately disappear beneath the pats digging and tunneling and, if they successfully adapt to their new environment, soon become a permanent, self-sustaining part of the local ecology. In time they multiply and within three or four years the benefits to the pasture are obvious.

Dung beetles work from the inside of the pat so they are sheltered from predators such as birds and foxes. Most species burrow into the soil and bury dung tunnels directly underneath the pats, which are hollowed out from within. Some large species originating from France excavate tunnels to a depth of approximately 30cm below the dung pat. These beetles make sausage-shaped brood chambers along the tunnels. The shallowest tunnels belong to a much smaller Spanish species that buries dung in chambers that hang like fruit from branches of a pear tree. South African beetles dig narrow tunnels of approximately 20 cm below the surface of the pat. Some surface-dwelling beetles, including a south African species, cut perfectly-shaped balls from the pat, which are rolled away and attached to the bases of the plants.

For maximum dung burial in spring, summer and autumn, farmers require a variety of species with overlapping periods of activity. In the cooler environments of

the state of Victoria, the large French species (2.5 cms long) is matched with smaller (half this size), temperate-climate Spanish species. The former are slow to recover from the winter cold and produce only one or two generations of offspring from late spring until autumn. The latter, which multiply rapidly in early spring, produce two to five generations annually. The south African ball-rolling species, being a sub-tropical beetle, prefers the climate of northern and coastal New South Wales where it commonly works with the South African tunneling species. In warmer climates, many species are active for longer periods of the year.

Dung beetles were initially introduced in the 1960s with a view to controlling buffalo flies by removing the dung within a day or two and so preventing flies from breeding. However, other benefits have become evident. Once the beetle larvae have finished pupation, the residue is a first-rate source of fertilizer. The tunnels abandoned by the beetles provide excellent aeration and water channels for root systems. In addition, when the new generation of beetles has left the nest the abandoned burrows are an attractive habitat for soil-enriching earthworms, which decompose it further to provide essential soil nutrients. If it were not for the dung beetle, chemical fertiliser and dung would be washed by rain into streams and rivers before it could be absorbed into the hard earth, polluting water courses and causing blooms of blue-green algae. Without the beetle to dispose of the dung, cow pats would litter pastures making grass inedible to cattle and depriving the soil of sunlight. Australia's 30 million cattle each produce 10-12 cow pats a day. This amounts to 1.7 billion tones a year, enough to smother about 11,000 sq km of pasture, half the area of Victoria.

Dung beetles have become an integral part of the successful management of dairy farms in Australia over the past few decades. A number of species are available from the CSIRO or through a small number of private breeders, most of whom were entomologists with the CSIRO's dung beetle unit who have taken their specialised knowledge of the insect and opened small businesses in direct competition with their former employer.

Glossary

1. dung: the droppings or excreta of animals
2. cow pats: droppings of cows

Question 9-13

Complete the table below.

Choose NO MORE THAN THREE WORDS OR A NUMBER from Reading Passage 1 for each answer.

Write your answers in boxes 9-13 on your answer sheet.

Species	Size	Preferred climate	Complementary species	Start of active period	Number of generations per year
French	2.5 cm	Cool	Spanish	late spring	1-2
Spanish	1.25 cm	9		10	11
South African ball roller		12	13		

Section A

The role of government in environmental management is difficult but inescapable. Sometimes, the state tries to manage the resources it owns, and does so badly. Often, however, governments act in an even more harmful way. They actually subsidize the exploitation and consumption of natural resources. A whole range of policies, from farm-price support to protection for coal-mining, do environmental damage and (often) make no economic sense. Scrapping them offers a two-fold bonus: a cleaner environment and a more efficient economy. Growth and environmentalism can actually go hand in hand, if politicians have the courage to confront the vested interest that subsidies create.

Section B

No activity affects more of the earth's surface than farming. It shapes a third of the planet's land area, not counting Antarctica, and the proportion is rising. World food output per head has risen by 4 per cent between the 1970s and 1980s mainly as a result of increases in yields from land already in cultivation, but also because more land has been brought under plough. Higher yields have been achieved by increased irrigation, better crop breeding, and a doubling in the use of pesticides and chemical fertilizers in the 1970s and 1980s.

Section C

All these activities may have damaging environmental impacts. For example, land clearing for agriculture is the largest single cause of deforestation; chemical fertilisers and pesticides may contaminate water suppliers; more intensive farming and the abandonment of fallow periods tend to exacerbate soil erosion; and the spread of monoculture and use of high-yielding varieties of food plants which might have provided some insurance against pests and diseases in future. Soil erosion threatens the productivity of land in both rich and poor countries. The United States, where the most careful measurements have been done, discovered in 1982 that about one-fifth of its farmland was losing topsoil at a rate likely to diminish the soil's productivity. The country subsequently embarked upon a program to convert 11 per cent of its cropped land to meadow or forest. Topsoil in India and China is vanishing much faster in America.

Section D

Government policies have frequently compounded the environmental damage that farming can cause. In the rich countries, subsidies for growing crops and price supports for farm output drive up the price of land. The annual value of these subsidies is

immense: about \$250 billion, or more than all World Bank lending in the 1980s. To increase the output of crops per acre, a farmer's easiest option is to use more of the most readily available inputs: fertilisers and pesticides. Fertiliser use doubled in Denmark in the period 1960-1985 and increased in the Netherlands by 150 per cent. The quantity of pesticides applied has risen too: by 69 per cent in 1975-1984 in Denmark, for example, with a rise of 115 per cent in the frequency of application in the three years from 1981.

In the late 1980s and early 1990s some efforts were made to reduce farm subsidies. The most dramatic example was that of New Zealand, which scrapped most farm support in 1984. A study of the environmental effects, conducted in 1993, found that the end of fertiliser subsidies had been followed by a fall in fertiliser use (a fall compounded by the decline in world commodity prices, which cut farm incomes). The removal of subsidies also stopped land-clearing and over-stocking, which in the past had been the principal causes of erosion. Farms began to diversify. The one kind of subsidy to manage soil erosion.

In less enlightened countries, and in the European Union, the trend has been to reduce rather than eliminate subsidies, and to introduce new payments to encourage farmers to treat their land in environmentally friendlier ways, or to leave it fallow. It may sound strange but such payments need to be higher than existing incentives for farmers to grow food crops. Farmers, however, dislike being paid to do nothing. In several countries they have become interested in the possibility of using fuel produced from crops residues either as a replacement for petrol (as ethanol) or as fuel for power stations (as biomass). Such fuels produce far less carbon dioxide than coal or oil, and absorb carbon dioxide as they grow. They are therefore less likely to contribute to the greenhouse effect. But they are rarely competitive with fossil fuels unless subsidized – and growing them does not less environmental harm than other crops.

Section E

In poor countries, governments aggravate other sorts of damage. Subsidies for pesticides and artificial fertilisers encourage farmers to use greater quantities than are needed to get the highest economic crop yield. A study by the International Rice Research Institute of pesticide use by farmers in South East Asia found that, with pest-resistant varieties of rice, even moderate applications of pesticide frequently cost farmers more than they saved. Such waste puts farmers on a chemical treadmill: bugs and weeds become resistant to poisons, so next year's poisons must be more lethal. One cost is to human health. Every year some 10,000 people die from pesticide poisoning, almost all of them in the developing countries, and another 400,000 become seriously ill. As for artificial fertilisers, their use world-wide increased by 40 per cent unit of farmed land between the mid 1970s and 1980, mostly in the developing countries. Overuse of fertilisers may cause farmers to stop rotating crops or leaving their land fallow. That, in turn, may make soil erosion worse.

Section F

A result of the Uruguay Round of world trade negotiations is likely to be a reduction of 36 per cent in the average levels of farm subsidies paid by the rich countries in the 1986-1990. Some of the world's food production will move from Western Europe to regions where subsidies are lower or non-existent, such as the former communist

countries and parts of the developing world. Some environmentalists worry about this outcome. It will undoubtedly mean more pressure to convert natural habitat into farmland. But it will also have many desirable environmental effects. The intensity of farming in the rich world should decline, and the use of chemical inputs will diminish. Crops are more likely to be grown in the environments to which they are naturally suited. And more farmers in poor countries will have the money and the incentive to manage their land in ways that are sustainable in the long run. That is important. To feed an increasingly hungry world, farmers need every incentive to use their soil and water effectively and efficiently.

Questions 19-22

Complete the table below using information in sections B and C of Reading Passage 2.

Choose your answers A-G from the box below the table and write them in boxes 19-22 on your answer sheet.

Agricultural practice	Environmental damage that may result
• 19	• Deforestation
• 20	• Degraded water supply
• More intensive farming	• 21
• Expansion of monoculture	• 22

- | |
|--|
| <p>A Abandonment of fallow period
 B Disappearance of old plant varieties
 C Increased use of chemical inputs
 D Increased irrigation
 E Insurance against pests and diseases
 F Soil erosion
 G Clearing land for cultivation</p> |
|--|

Question 28

From the list below choose the most suitable title for Reading Passage 2.

Write the appropriate letter A-E in the box 28 on your answer sheet.

- A Environmental management
B Increasing the world's food supply
C Soil erosion
D Fertilisers and pesticides – the way forward
E Farm subsidies

Reading passage 3

You should spend about 20 minutes on Question 29-40 which are based on Reading Passage 3 below.

THE CONCEPT OF ROLE THEORY

Role set

Any individual in any situation occupies a role in relation to other people. The particular individual with whom is one concerned in the analysis of any situation is usually given the name of *focal person*. He has the *focal role* and can be regarded as sitting in the middle of a group of people, with whom he interacts in some way in that situation. This group of people is called his *role set*. For instance, in the family situation, an individual's role set might be shown in *Figure 6*.

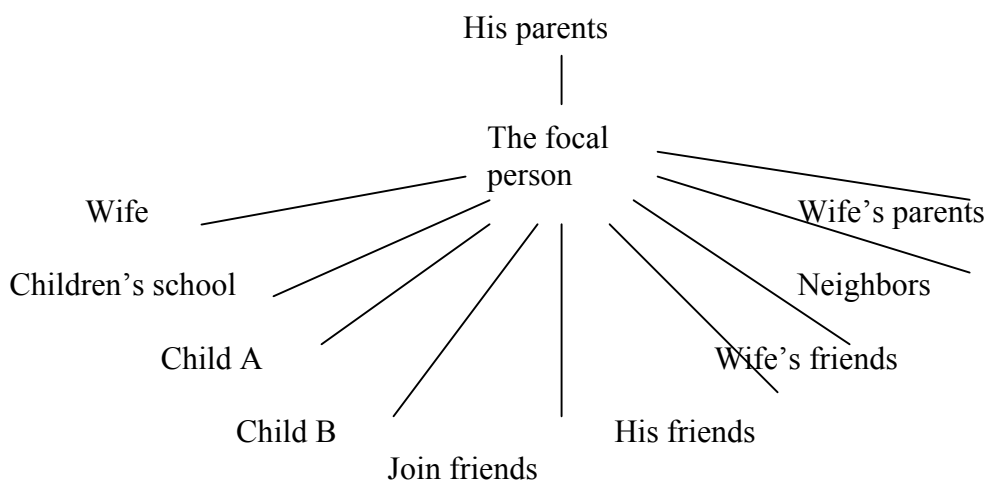


Figure 6

The role set should include all those with whom the individual has more than trivial interactions.

Role definition

The definition of any individual's role in any situation will be a combination of the *role expectations* that the members of the role set have of the focal role. These expectations are often occupationally defined, sometimes even legally so. The role definitions of lawyers and doctors are fairly clearly defined both in legal and in cultural terms. The role definitions of, say, a film star or bank manager, are also fairly clearly perhaps. Individuals often find it hard to escape from the role that cultural traditions have defined for them. Not only with doctors or lawyers is the required role behaviour so constrained that if you are in that role for long it eventually becomes part of *you*, part of your personality. Hence, there is *some* likelihood that all accountants will be alike or that all blondes are similar – they are forced that way by the expectations of their role.

It is often important that you make it clear what your particular role is at a given time. The means of doing this are called, rather obviously, *role signs*. The simplest of role signs is a uniform. The number of stripes on your arm or pips on your shoulder is a

very precise role definition which allows you to do certain very prescribed things in certain situations. Imagine yourself questioning a stranger on a dark street at midnight without wearing the role signs of a policeman!

In social circumstances, dress has often been used as a role sign to indicate the nature and degree of formality of any gathering and occasionally the social status of people present. The current trend towards blurring these role signs is probably democratic, but it also makes some people very insecure. Without role signs, who is to know who has what role?

Place is another role sign. Managers often behave very differently outside the office and in it, even to the same person. They use a change of location to indicate a change in role from, say, boss to friend. Indeed, if you want to change your roles you must find some outward sign that you are doing so or you won't be permitted to change – the subordinate will continue to hear you as his boss no matter how hard you try to be his friend. In very significant cases of role change, e.g. from a soldier in the ranks to officer, from bachelor to married man, the change of role has to have a very obvious *sign*, hence *rituals*. It is interesting to observe, for instance, some decline in the emphasis given to marriage rituals. This could be taken as an indication that there is no longer such a big change in role from single to married person, and therefore no need for a public change in *sign*.

In organisations, office signs and furniture are often used as role signs. These and other perquisites of status are often frowned upon, but they may serve a purpose as a kind of uniform in a democratic society; roles without signs often lead to confused or differing expectations of the role of the focal person.

Role ambiguity

Role ambiguity results when there is some uncertainty in the minds, either of the focal person or of the members of his role set, as to precisely what his role is at any given time. One of the crucial expectations that shape the role definition is that of the individual, the focal person himself. If his occupation of the role is unclear, or if it differs from that of the others in the role set, there will be a degree of role ambiguity. Is this bad? Not necessarily, for the ability to shape one's own role is one of the freedoms that many people desire, but the ambiguity may lead to role stress which will be discussed later on. The virtue of the job description is that they lessen this role ambiguity. Unfortunately, job descriptions are seldom complete role definitions, except at the lower end of the scale. At middle and higher management levels, they are often a list of formal jobs and duties that say little about more subtle and informal expectations of the role. The result is therefore to give the individual an uncomfortable feeling that there are things left unsaid, i.e. to *heighten* the sense of role ambiguity.

Looking at role ambiguity from the other side, from the point of view of the members of the role set, lack of clarity in the role of the focal person can cause insecurity, lack of confidence, irritation and even anger among members of his role set. One list of the roles of a manager identified the following: executive, planner, policy maker, expert, controller of rewards and punishments, counselor, friend, teacher. If it is not clear, through role signs of one sort or another, which role is currently the operational one, the other party may not react in the appropriate way – we may, in fact, hear quite

another message if the local person speaks to us, for example, as a teacher and we hear her as an executive.

Questions 36-39

Choose **ONE OR TWO WORDS** from Reading Passage 3 for each answer.

Write your answers in boxes 36-39 on your answer sheet.

- 36** A new headmaster of a school who enlarges his office and puts in expensive carpeting is using the office as a ...
- 37** The graduation ceremony in many universities is an important ...
- 38** The wig which judges wear in UK courts is a ...
- 39** The parents of students in a school are part of the headmaster's...

APPENDIX 3 IELTS Examination – test 4

Reading passage 3

You should spend about 20 minutes on **Question 28-40** which are based on Reading Passage 3 below.

Measuring Organisational Performance

There is clear-cut evidence that, for a period of at least one year, supervision which increases the direct pressure for productivity can achieve significant increases in production. However, such short-term increases are obtained only at a substantial and serious cost to the organization.

To what extent can a manager make an impressive earnings record over a short period of one to three years by exploiting the company's investment in the human organisation in his plant or division? To what extent will the quality of his organisation suffer if he does so? The following is a description of an important study conducted by the Institute for Social Research designed to answer these questions.

The study covered 500 clerical employees in four parallel divisions. Each division was organised in exactly the same way, used the same technology, did exactly the same kind of work, and had employees of comparable aptitudes.

Productivity in all four of the divisions depended on the number of clerks involved. The work entailed the processing of accounts and generating of invoices. Although the volume of work was considerable, the nature of the business was such that it could only be processed as it came along. Consequently, the only way in which productivity could be increased was to change the size of the work group.

The four divisions were assigned to two experimental programmes on a random basis. Each programme was assigned at random a division that had been historically high in productivity and a division that had been below average in productivity. No attempt was made to place a division in the programme that would best fit its habitual methods of supervision used by the manager, assistant managers, supervisors and assistant supervisors.

The experiment at the clerical level lasted for one year. Beforehand, several months were devoted to planning, and there was also a training period of approximately six months. Productivity was measured continuously and computed weekly throughout the year. The attitudes of employees and supervisory staff towards their work were measured just before and after the period.

Turning now to the heart of the study, in two divisions an attempt was made to change the supervision so that the decision levels were pushed *down* and detailed supervision

of the workers reduced. More general supervision of the clerks and their supervisors was introduced. In addition, the managers, assistant managers, supervisors and assistant supervisors of these two divisions were trained in group methods of leadership, which they endeavored to use as much as their skill would permit during the experimental year. For easy reference, the experimental changes in these two divisions will be labeled the 'participative programme'.

In the other two divisions, by contrast, the programme called for modifying the supervision so as to increase the closeness of supervision and move the decision levels *upwards*. This will be labeled the 'hierarchically controlled programme'. These changes were accomplished by a further extension of the scientific management approach. For example, one of the major changes made was to have the jobs timed and to have standard times computed. This showed that these divisions were overstaffed by about 30%. The general manager then ordered the managers of these two divisions to cut staff by 25%. This was done by transfers without replacing the persons who left; no one was to be dismissed.

Results of the Experiments

Changes in Productivity

Figure 1 shows the changes in salary costs per unit of work, which reflect the change in productivity that occurred in the divisions. As will be observed, the hierarchically controlled programmes increased productivity by about 25%. This was a result of the direct orders from the general manager to reduce staff by that amount. Direct pressure produced a substantial increase in production.

A significant increase in productivity of 20% was also achieved in the participative programme, but this was not as great as increase as in the hierarchically controlled programme. To bring about this improvement, the clerks themselves participated in the decision to reduce the size of the work group. (They were aware of course that productivity increases were sought by management in conducting experiments.) Obviously, deciding to reduce the size of a work group by eliminating some of its members is probably one of the most difficult decisions for a work group to make. Yet the clerks made it. In fact, one division in the participative programme increased its productivity by about the same amount as each of the two divisions in the hierarchically controlled programme. The other participative division, which historically had been the poorest of all the divisions, did not do so well and increased productivity by only 15%.

Changes in Attitudes

Although both programmes had similar effects on productivity, they had significantly different results in other respects. The productivity increases in the hierarchically controlled programme were accompanied by shifts in an adverse direction in such factors as loyalty, attitudes, interest, and involvement in the work. But just the opposite was true in the participative programme.

For example, Figure 2 shows that when more general supervision and increased participation were provided, the employees' feeling of responsibility to see that the work got done increased. Again, when the supervisor was away, they kept on working. In the hierarchically controlled programme, however, the feeling of responsibility decreased, and when the supervisor was absent, work tended to stop.

As Figure 3 shows, the employees in the participative programme at the end of the year felt that their manager and assistant manager were 'closer to them' than at the beginning of the year. The opposite was true in the hierarchically controlled programme. Moreover, as Figure 4 shows, employees in the participative programme felt that their supervisors were more likely to 'pull' for them, or for the company and them, and not be solely interested in the company, while in the hierarchically controlled programme, the opposite trend occurred.

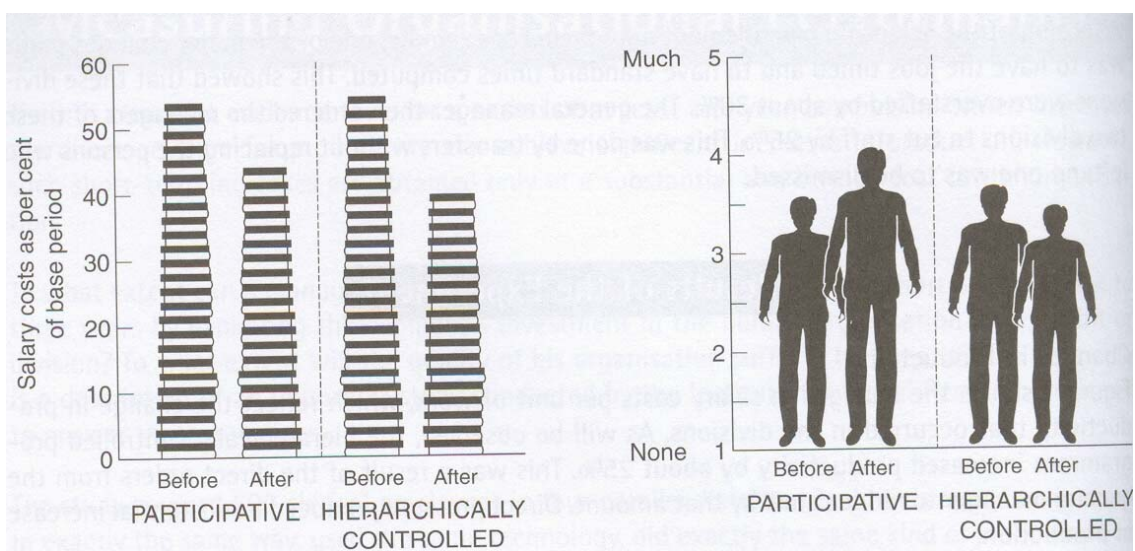


FIGURE 1

FIGURE 2

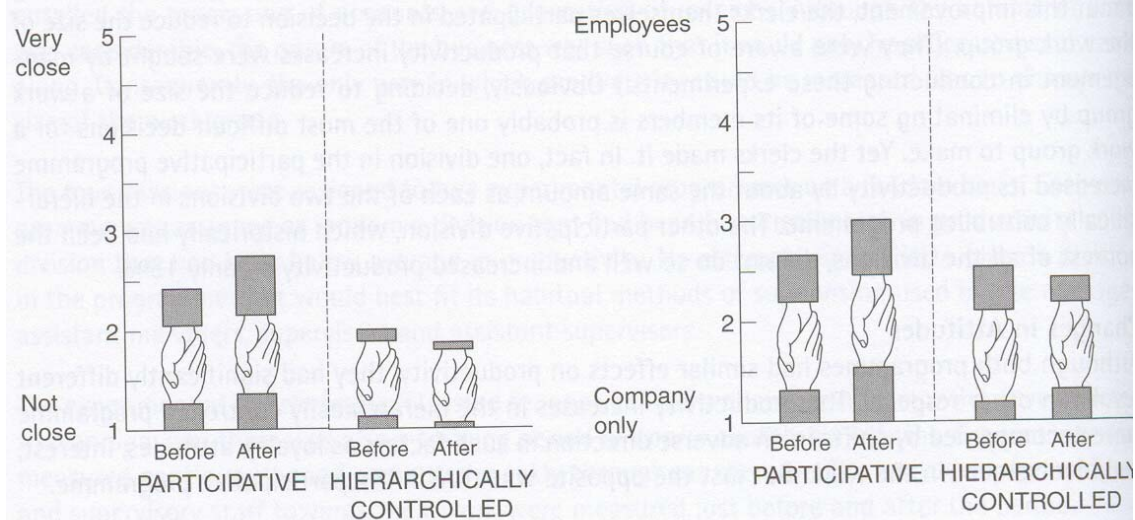


FIGURE 3

FIGURE 4

Questions 31-36

Complete the summary below. Choose **ONE** word from Reading Passage 3 for each answer.

Write your answers in boxes 31-36 on your answer sheet.

This experiment involved an organisation comprising four divisions, which were divided into two programmes: the hierarchically controlled programme and the participative programme. For a period of one year a different method of ... **31** ... was used in each programme. Throughout this time ... **32** ... was calculated on a weekly basis. During the course of the experiment the following changes were made in an attempt to improve performance.

In the participative programme:

- supervision of all workers was ... **33** ...
- supervisory staff were given training in ... **34** ...

In the hierarchically controlled programme:

- supervision of all workers was increased.
- work groups were found to be ... **35** ... by 30%.
- the work force was ... **36** ... by 25%.

APPENDIX 4 TOEFL Test

Opportunists and Competitors

➔ Growth, reproduction, and daily metabolism all require an organism to expend energy. The expenditure of energy is essentially a process of budgeting, just as finances are budgeted. If all of one's money is spent on clothes, there may be none left to buy food or go to the movies. Similarly, a plant or animal cannot squander all its energy on growing a big body if none would be left over for reproduction, for this is the surest way to extinction.

All organisms, therefore, allocate energy to growth, reproduction, maintenance, and storage. No choice is involved; this allocation comes as part of the genetic package from the parents. Maintenance for a given body design of an organism is relatively constant. Storage is important, but ultimately that energy will be used for maintenance, reproduction, or growth. Therefore the principal differences in energy allocation are likely to be between growth and reproduction.

Almost all of an organism's energy can be diverted to reproduction, with very little allocated to building the body. Organisms at this extreme are "opportunists." At the other extreme are "competitors," almost all of whose resources are invested in building a huge body, with a bare minimum allocated to reproduction.

Dandelions are good examples of opportunists. Their seedheads raised just high enough above the ground to catch the wind, the plants are no bigger than they need be, their stems are hollow, and all the rigidity comes from their water content. Thus, a minimum investment has been made in the body that becomes a platform for seed dispersal. These very short-lived plants reproduce prolifically; that is to say they provide a constant rain of seed in the neighborhood of parent plants. A new plant will spring up wherever a seed falls on a suitable soil surface, but because they do not build big bodies, they cannot compete with other plants for space, water, or sunlight. These plants are termed opportunists because they rely on their seeds' falling into settings where competing plants have been removed by natural processes, such as along an eroding riverbank, on landslips, or where a tree falls and creates a gap in the forest canopy.

Opportunists must constantly invade new areas to compensate for being displaced by more competitive species. Human landscapes of lawns, fields, or flowerbeds provide settings with bare soil and a lack of competitors that are perfect habitats for colonization by opportunists. ■ Hence, many of the strongly opportunistic plants are the common weeds of fields and gardens. ■

Because each individual is short-lived, the population of an opportunist species is likely to be adversely affected by drought, bad winters, or floods. ■ If their population is tracked through time, it will be seen to be particularly unstable—soaring and plummeting in irregular cycles. ■

➡ The opposite of an opportunist is a competitor. These organisms tend to have big bodies, are long-lived, and spend relatively little effort each year on reproduction. An oak tree is a good example of a competitor. A massive oak claims its ground for 200 years or more, outcompeting all other would-be canopy trees by casting a dense shade and drawing up any free water in the soil. The leaves of an oak tree taste foul because they are rich in tannins, a chemical that renders them distasteful or indigestible to many organisms. The tannins are part of the defense mechanism that is essential to longevity. Although oaks produce thousands of acorns, the investment in a crop of acorns is small compared with the energy spent on building leaves, trunk, and roots. Once an oak tree becomes established, it is likely to survive minor cycles of drought and even fire. A population of oaks is likely to be relatively stable through time, and its survival is likely to depend more on its ability to withstand the pressures of competition or predation than on its ability to take advantage of chance events. It should be noted, however, that the pure opportunist or pure competitor is rare in nature, as most species fall between the extremes of a continuum, exhibiting a blend of some opportunistic and some competitive characteristics.

1. The word squander in the passage is closest in meaning to

- extend
- transform
- activate
- waste

2. The word none in the passage refers to

- food
- plant or animal
- energy
- big body

3. In paragraph 1, the author explains the concept of energy expenditure by

- identifying types of organisms that became extinct
- comparing the scientific concept to a familiar human experience
- arguing that most organisms conserve rather than expend energy
- describing the processes of growth, reproduction, and metabolism

Paragraph 1 is marked with an arrow [➡].

4. According to the passage, the classification of organisms as “opportunists” or “competitors” is determined by

- how the genetic information of an organism is stored and maintained
- the way in which the organism invests its energy resources

- whether the climate in which the organism lives is mild or extreme
- the variety of natural resources the organism consumes in its environment

5. The word **dispersal** in the passage is closest in meaning to

- development
- growth
- distribution
- protection

6. Which of the sentences below best expresses the essential information in the highlighted sentence in the passage? *Incorrect* choices change the meaning in important ways or leave out essential information.

- Because their seeds grow in places where competing plants are no longer present, dandelions are classified as opportunists.
- Dandelions are called opportunists because they contribute to the natural processes of erosion and the creation of gaps in the forest canopy.
- The term opportunists applies to plants whose seeds fall in places where they can compete with the seeds of other plants.
- The term opportunists applies to plants whose falling seeds are removed by natural processes.

7. The word **massive** in the passage is closest in meaning to

- huge
- ancient
- common
- successful

8. All of the following are mentioned in paragraph 7 as contributing to the longevity of an oak tree EXCEPT

- the capacity to create shade
- leaves containing tannin
- the ability to withstand mild droughts and fire
- the large number of acorns the tree produces

Paragraph 7 is marked with an arrow [➔].

9. According to the passage, oak trees are considered competitors because

- they grow in areas free of opportunists

- they spend more energy on their leaves, trunks and roots than on their acorns
- their population tends to increase or decrease in irregular cycles
- unlike other organisms, they do not need much water or sunlight

10. In paragraph 7, the author suggests that most species of organisms

- are primarily opportunists
- are primarily competitors
- begin as opportunists and evolve into competitors
- have some characteristics of opportunists and some of competitors

Paragraph 7 is marked with an arrow [➔].

11. Look at the four squares [■] that indicate where the following sentence could be added to the passage.

Such episodic events will cause a population of dandelions, for example, to vary widely.

Where would the sentence best fit?

Click on a square [■] to add the sentence to the passage.

12. Directions: Complete the table by matching the phrases below

Directions: Select the appropriate phrases from the answer choices and match them to the type of organism to which they relate. TWO of the answer choices will NOT be used. *This question is worth 4 points.*

Drag your answer choices to the spaces where they belong. To remove an answer choice, click on it. To review the passage, click on **View Text**.

Answer Choices

Opportunists

Vary frequently the amount of energy they spend in body maintenance



Have mechanisms for protecting themselves from predation



...



organisms have been removed

Have relatively short life spans ●

Invest energy in the growth of large, strong structures ●

Have populations that are unstable in response to climate conditions ●

Can rarely find suitable soil for reproduction ●

Produce individuals that can withstand changes in the environmental conditions ●

Reproduce in large numbers ●

Competitors

APPENDIX 5
TOEFL Test – item 13

Line
(5) Barbara Kasten is an artist who makes photographs for constructions that she creates for the purpose of photographing them. In her studio she arranges objects such as mirrors, solid forms, and flat surfaces into what could be called large still life arrangements, big enough to walk into. She lights the construction, then rearranges and photographs it until she arrives at a final image. She also photographs away from her studio at various architectural sites, bringing camera, lights, mirrors, and a crew of assistants to transform the site into her own abstract image.

(10) Kasten starts a studio construction with a simple problem, such as using several circular and rectangular mirrors. She puts the first objects in place, sets up a camera, then goes back and forth arranging objects and seeing how they appear in the camera. Eventually she makes instant color prints to see what the image looks like. At first she works only with objects, concentrating on their composition; then she lights them and adds color from lights covered with colored filters.

(15) Away from the studio, at architectural sites, the cost of the crew and equipment rental means she has to know in advance what she wants to do. She visits each location several times to make sketches and test shots. Until she brings in the lights, however, she cannot predict exactly what they will do to the image, so there is some improvising on the spot.

15. The word “transform” in line 6 is closest in meaning to
- (A) move
 - (B) extend
 - (C) change
 - (D) interpret

APPENDIX 6
TOEFL Test – item 14

The temperature of the Sun is over 5,000 degrees Fahrenheit at the surface, but it rises to perhaps more than 16 million degrees at the center. The Sun is so much hotter than the Earth that matter can exist only as a gas, except at the core. In the
Line core of the Sun, the pressures are so great against the gases that, despite the high
(5) temperature, there may be a small solid core. However, no one really knows, since the center of the Sun can never be directly observed.

Solar astronomers do know that the Sun, the zones are the corona, chromosphere, photosphere, convection zone, and finally the core. The first three zones are regarded as the Sun's atmosphere. But since the Sun has no solid
(10) surface, it is hard to tell where the atmosphere ends and the main body of the Sun begins.

The Sun's outermost layer begins about 10,000 miles above the visible surface and goes outward for millions of miles. This is the only part of the Sun that can
(15) be seen only when special instruments are used on cameras and telescopes to shut out the glare of the Sun's rays.

The corona is a brilliant, pearly white, filmy light, about as bright as the full Moon. Its beautiful rays are a sensational sight during an eclipse. The corona's rays flash out in a brilliant fan that has wispy spikelike rays near the Sun's north
(20) and south poles. The corona is thickest at the Sun's equator.

The corona rays are made up of gases streaming outward at tremendous speeds and reaching a temperature of more than 2 million degrees Fahrenheit. The rays of gas thin out as they reach the space around the planets. By the time the Sun's corona rays reach the Earth, they are weak and invisible.

24. The word "great" in line 4 is closest in meaning to
- (A) dangerous
 - (B) unknown
 - (C) variable
 - (D) strong

APPENDIX 7
TOEFL Test – item 15

Read the passage provided and select the correct answer choice for each question.

TOEFL – READING COMPREHENSION

One of the most renowned Spanish architects of all time was Antoni Gaudi. Gaudi's emergence as one of Spain's preeminent artists at the end of the nineteenth century marked a milestone in the art world.

Gaudi's popularity helped to bring about the acceptance and rebirth of the Catalan language, which had been banned during the peak of Castilian literature and art. Gaudi shares his Catalonian background with two other famous Spanish artists, Pablo Picasso and Miro. The diverse ethnic background of the region greatly influenced the work of Picasso and Miro, as well as Gaudi. Thus, their works were a combination of an old history and an active, vivid imaginary world. This has sometimes been referred to as the "Catalan Mind." Yet it was perhaps Gaudi who had the greatest talent for bringing together diverse groups, ones which others viewed as being too diametrically opposed to be capable of coming together and co-existing amicably.

This was apparent not only in the artist artists and other individuals who surrounded him, but also in the varied styles and techniques he employed in his architecture. Much of his work can be seen in Barcelona, where his structures are known as a fine representation of modernism. He also used a great variety of color in his buildings, and this art nouveau is often associated with his own unique style of deign.

All of these factors are what helped put him at the forefront of art movements to come: his unique ability to take on and transform traditional Spanish elements with the emerging diverse ethnic groups, merging these with his own fertile imagination, and consequently turning these forces into some of the greatest architecture the world has ever seen.

1. Antoni Gaudi's fame is due primarily to his world-famous
 - (A) paintings
 - (B) architectural structures
 - (C) political skills
 - (D) business acumen

APPENDIX 8

EAP Teacher 1

I - Look at the sentences below. All the words in italics are nonsense words. Discover what these words mean from the context of the sentence. Sometimes more than one word is possible.

- 1 – It was a very cold day, so I put a *tribbet* around my neck.
- 2 – He was so *fliglive* that he drank a whole bottle of Coke.
- 3 – Mary did three tralets yesterday but failed them all because she hadn't studied enough
- 4 - She did the exam very *trodly* because she had a headache.
- 5 – The doctor *sarked* very late at work because he overslept.

II - Complete the gaps in the text with the correct words

Americans are well-known for being If we're taking a in the park and we pass someone, we usually say *hi!* or *how's it going?* to each And we usually say a few words to people in stores, bars, and banks. But remember: is not friendship: it's In the United States, it's just as to make real friends as it is anywhere else.

hard – politeness – friendly – walk – other – friendliness

III - What are the articles related to these headlines?

- 1 – Teenagers say AIDS is their biggest fear
- 2 – World champion swimmer suspended after drug test
- 3 – Explosion kills 20 people

IV - Match the second part of each sentence

- | | |
|--|---|
| 1 – I speak fluent German, | () but I enjoy dancing. |
| 2 – we aren't going to Germany, | () and knows many good restaurants. |
| 3 – I don't do any sports, | () so there's no need to buy tickets. |
| 4 – Steven eats out a lot with friends | () but there were some strange people in the restaurant. |
| 5 – At first, everything seemed fine, | () and I've just come back from Germany. |

V - Read the text below and give three reasons why flying is bad for people's health

VI - Explain the following compound nouns from the text: heart attack, economy class, leg room, time zone, blood pressure.

Text:

Warning: Flying Is Bad For Your Health

Flying is the safest way to travel...or is it? Some doctors think the airplane is a dangerous place, especially for the old or the unhealthy. Although the airplane is pressurized, there is less oxygen than on the ground. So anyone who has had a heart attack should not fly for at least two weeks after the attack. After an operation, you should stay on the ground for at least ten days. Sitting on a plane for many hours - especially in economy class- gives everyone aches and pains, so you should get some exercises, especially on long flights.

Flying also causes dehydration. If you drink or eat too much, you'll wake up feeling sick. Everyone needs to drink more in the air, but you shouldn't drink alcohol because it makes you even more thirsty. The most common problem is jet lag. You should change to your new time zone as soon as possible, and you shouldn't sleep if it's still daylight. Crowded airports, long lines, and delays cause stress and high blood pressure. So, be careful! Flying is the safest way to travel, but is it the healthiest?

APPENDIX 9

EAP Teacher 2

Luisa May Alcott, an American author best known for her children's books *Little Women*, *Little Men*, and *Jo's Boys*, was profoundly influenced by her family, particularly her father. She was the daughter of Bronson Alcott, a well-known teacher, intellectual, and free thinker who advocated abolitionism, women's rights, and vegetarianism long before they were popular. He was called a man of unparalleled intellect by his friend Ralph Waldo Emerson. Bronson Alcott instilled in his daughter his lofty and spiritual values and in return was idolized by his daughter.

The financial situation of the Alcott family during Luisa's childhood was not good, mainly due to the fact that her father made unsound investments in projects that reflected his idealistic view of the world. As a result, Luisa had to begin helping to support her family at a young age, by taking a variety of low-paying jobs as a seamstress, a maid, and a tutor.

Her novel *Little Women* was patterned after her own family, and Louisa used her father as a model for the impractical yet serenely wise and adored father in *Little Women*. With the success of this novel she was able to provide for her family, giving her father financial security that until then he had never experienced.

2 - Na linha 2, a palavra “particularly” assemelha-se mais em termos de significado a (the word ‘particularly’ in line 2 is closest in meaning to):

- e) parcialmente por (partially for)
- f) estranhamente (strangely)
- g) exceto por (except for)
- h) especialmente (particularly)

4 - Na linha 5, a palavra “lofty” assemelha-se mais em termos de significado a (the word “lofty” in line 5 is closest in meaning to):

- e) comum (common, ordinary)
- f) generoso (generous)
- g) egoísta (selfish)
- h) simpático (nice)

6 - Qual das seguintes atividades NÃO foi exercida por Luisa para ganhar dinheiro em sua juventude (which of the following jobs did Luisa NOT take to earn her living at a very young age)?

- e) trabalhou como costureira (worked as a seamstress)
- f) trabalhou como faxineira (worked as a maid)
- g) trabalhou como professora (worked as a tutor, teacher)
- h) trabalhou em uma loja (worked at a store)

8 - Pode-se inferir a partir do texto que Luisa May Alcott usou o sucesso de *Little Women* para (It is possible to infer from the text that Luisa May Alcott used the success of *Little Women* to)

- e) presentear-se com tudo o que sempre quis (to buy herself all the presents she always wanted)
- f) atingir sucesso financeiro e pessoal (attain personal and financial success)
- g) dar a seu pai uma prova intangível de seu amor (give her father intangible proof of her love)
- h) separar-se de sua família (separate from her family)

9 - O propósito do autor nessa passagem é (the purpose of the author in the passage is):

- e) explicar como a autora tornou-se famosa (explain how the writer became famous)
- f) descrever a influencia da família na vida da escritora (describe the influence of the family on the life of the writer)
- g) apoiar as teorias educacionais de Bronson Alcott (support the educational theories by Bronson Alcott)
- h) mostrar o sucesso que pode ser atingido por um(a) autor(a) (show the success a writer can achieve)

11 - Retire do texto palavras formadas por prefixação e duas formadas por sufixação e suas respectivas paráfrases. (Identify in the text two words with suffixes and two words with prefixes and their respective paraphrases).

12 - Retire do texto uma conjunção de resultado e uma conjunção de adição (Identify in the text a causal conjunction and an additive conjunction).

13 - Retire do texto quatro grupos nominais com suas respectivas paráfrases (Identify in the text four nominal groups with their respective paraphrases).

14 - Diga a que se refere os seguintes referentes contextuais (What do the following referents refer to?).

15 - Traduza o segundo paragrafo do texto (Translate the second paragraph of the text).

APPENDIX 10

**Correspondence of the test tasks as presented within UFSC examinations
compared to the tasks required for university studies**

Analysis of the correspondence between the characteristics of TLU situation tasks and test tasks of UFSC examination

Characteristics of the Test tasks		Correspondence
1. Characteristics of the setting		Always low
2. Characteristics of the test rubrics		Always low
3. Characteristics of the input	Input may refer to both the texts and the tasks presented to the test taker	
<i>3.1 Format</i>		
Form	Language, and non-language (pictures) as accessory	High
Language	Both the tasks and the texts are in the target language, English	Tasks = low Texts = high
Length	Mostly short	Low
Type (text and task)	Text type: advertisements, texts adapted from texts published in coursebooks and magazines for teaching of English, dialogs.	Low
	Task type: Items	Low

To continue...

Table XX: analysis of the correspondence between the characteristics of TLU situation tasks and test tasks of UFSC examination (cont.)

	Characteristics of the Test tasks	Correspondence
<i>3.2 Language of input</i>	Input is to consider both the texts and the tasks presented to the test taker	
a) Language characteristics		
▪ Organizational characteristics: Grammatical	General and topic-related vocabulary Simple and complex syntax	High
▪ Organizational characteristics: Textual	Cohesion: Cohesion in some texts presented is different from the cohesion in academic expository texts, such as in the advertisements Rhetorical organization: - description of factual information (Charlie Chaplin) - problem/solution (Movies and Photographs)	Medium High
▪ Pragmatic characteristics: Functional	Ideational Manipulative (instrumental and regulatory)	High Medium
▪ Pragmatic characteristics: Sociolinguistic	Register – many texts with language more informal than academic texts Cultural reference – texts with cultural reference Figurative language – very little, if any	Low Medium High
b) Topical characteristics	Topics: personal, cultural, curiosities, variety	Low

To continue...

Table XX: analysis of the correspondence between the characteristics of TLU situation tasks and test tasks of UFSC examination (cont.)

	Characteristics of the Test tasks	Correspondence
4. Characteristics of the expected response		
<i>4.1 Format</i>		
Form	Language	High
Language	Target language = English	Low
Type of response	Selected receptive response and selected production response Set of binary item, where each and all proposition or phrase or sentence in a set can be either true or false as a response	Low
5. Relationship between input and response.		
<i>5.2 Scope of relationship</i>	Narrow scope, since some questions may be answered based on limited part of the text. Broad scope, since some questions must be answered based on longer discourse.	High
<i>5.3 Directness of relationship</i>	Direct and indirect – some responses may be given using primarily the information provided by the input, and may not.	High

Based on the following information about each text:

Text 1

An expository text, titled Charlie Chaplin – a comic genius, presenting description of factual information about Charlie Chaplin's early years of life, which is an adapted version of a text published on the 1999 August issue of the Brazilian Speak Up magazine.

- Language of the input
 - o Vocabulary: general and specific related to cinema and movies
 - o Syntax: simple and complex structures
 - o Functional: mainly ideational
 - o Sociolinguistic: middle degree of formality
 - o Topic: cinema and movie

Text 2

An expository text, titled Charlie Chaplin: the later years, presenting description of factual information about Charlie Chaplin's later years of life, which is an adapted version of a text published on the 1999 August issue of the Brazilian Speak Up magazine

- Language of the input
 - o Vocabulary: general and specific related to cinema and movies
 - o Syntax: simple and complex structures
 - o Functional: mainly ideational
 - o Sociolinguistic: middle degree of formality and some cultural reference
 - o Topic: cinema and movie

Text 3

An expository text, titled Rewards for talents, presenting description of factual information about awards for artists, published on Compact English Book, 1998.

- Language of the input
 - o Vocabulary: general and specific related to cinema and movies
 - o Syntax: simple and complex structures
 - o Functional: mainly ideational
 - o Sociolinguistic: middle degree of formality and some cultural reference
 - o Topic: cinema and movie

Text 4

An expository text, titled Movies and Photographs, presenting an invitation for traveling to learn about different cultures, published on the Lingua Inglesa: Leitura. Cortez, 1991.

- Language of the input
 - o Vocabulary: general and specific related to cinema and movies
 - o Syntax: mainly simple structures
 - o Functional: mainly ideational and manipulative
 - o Sociolinguistic: informal register
 - o Topic: cinema and movie

APPENDIX 11
UFSC 2003 Entrance Examination

TEXT 1

CHARLIE CHAPLIN – A COMIC GENIUS



One of the most important and influential figures in the history of motion pictures, Charlie Chaplin was perhaps the greatest comedian to have ever lived. He made his reputation in 1914 when, in his second film, *Kid Auto Races at Venice*, he introduced the world to the helpless “little tramp.” With his smudge moustache, baggy trousers and bowler hat, and twirling his cane, the tramp soon had cinema audiences entranced. It was a fantastic creation, stirring up emotions, both happy and sad, and Chaplin played that classic role in more than 70 films during his career, earning him both a fortune and international fame.

Chaplin’s beginnings never promised such success. Though born into a wealthy London family, the good times quickly disappeared. His father deserted when Charlie was an infant (and later died of alcoholism) and his mother, a successful music hall star, had a nervous breakdown and was sent to an asylum. Charlie thus found himself in an orphanage. It was the theatre that gave Chaplin his first release from the pressures of troubled life. He made his debut in 1894, appearing on stage with his mother. Later he became part of Fred Karno’s music hall troupe and went with them on their American tour of 1912. It was while the company was in the United States that the young Chaplin was spotted by the film director Mack Sennett and signed to Keystone Films at 150 dollars a week. Over the next few months Chaplin made dozens of films for Keystone many of which featured his newly created “little tramp” character.

From: *Speak Up*. Agosto 1999 – nº 147 (adapted).

1 - Read the summaries below. Which one(s) contains (contain) the same information found in the text?

01. According to the text, Chaplin stands as one of the greatest comedians ever, being also a relevant and powerful person in the history of the movie industry. His success is due to a character he created, known as the “little tramp”. First introduced to the world in *Kid Auto Races at Venice*, the “little tramp” appeared in all Chaplin’s movies and earned money and fame. Chaplin was meant to be successful since the beginning of his career. Born into a rich family, Chaplin was sent to an orphanage when his father and mother died. In 1912 he went on a tour with Karno’s music hall troupe, but his first performance on stage was in 1894. When touring with Karno’s group, Chaplin was invited to film Keystone for 150 dollars a week.
02. In the text it is said that Chaplin gained one of the highest positions as a comedian in the cinema world. The text also describes the character that brought Chaplin

fame and fortune and shows when his career blossomed. Besides that, we are told about who was responsible for his recognition and the number of times Chaplin performed his “little tramp” character in the films he took part in during those years. On the other hand, we learn how difficult Chaplin’s life was when he was very young.

04. The text refers to Charlie Chaplin as one of the greatest comic actors in the whole history of motion pictures. It also tells how Chaplin gained success through the creation of his famous character – the “little tramp” – and presents a brief description of him. Besides that, the reader is informed about the hard times Chaplin had to overcome still as a child, since his father left and his mother became seriously ill. The text also mentions when Chaplin’s talent was recognized and who took part in this process. Finally, the last lines of the text show us that Chaplin played the role of his new character in many films he made at that time.
08. Chaplin, the greatest comedian in the history of motion pictures, started his career in 1914, with *Kid Auto Races at Venice*. After having lived in an orphanage, he made his first public appearance in 1894, with his mother. Chaplin was invited to work for Keystone Films by Mack Sennet, who brought him fame and fortune.

2 - Choose the proposition(s) in which the definitions of the underlined words correspond to the meaning used in the text.

01. figures – numbered drawings or diagrams in a book.
02. cane – to punish someone, especially a child, by hitting them with a long thin stick.
04. role – the character played by an actor in a play or film.
08. stage – the raised floor in a theater on which plays are performed.
16. dozens of – a lot of.
32. featured – showed.

3 - Select the proposition(s) in which the beginning of the sentence can be correctly matched with both alternatives, according to the text.

01. With his “little tramp” character Chaplin
 a) received a large amount of money.
 b) became famous all over the world.
02. Charlie Chaplin’s beginnings were not easy because
 a) his family had serious problems.
 b) his father abandoned him and his mother got a mental illness when he was just a little boy.

04. As a “little tramp” Chaplin used to wear
 a) loose trousers.
 b) a hat with a round hard top.
08. In 1912 Chaplin
 a) traveled with a music hall company around the United States.
 b) made a show with his mother.
16. Every month Chaplin
 a) received almost two hundred dollars.
 b) was invited to make a new film.

TEXT 2

CHARLIE CHAPLIN: THE LATER YEARS

Chaplin's subsequent films, like *The Tramp* and *Shanghaied*, firmly established his reputation and, as his fame rose, so too did his salary and his power. By 1917 Chaplin was able to demand a million dollars for eight pictures. By now Chaplin was taking an increasing amount of control over his work: writing, directing, producing and even composing the music for many of the films in which he starred. In 1919 that control became complete with Chaplin, along with Mary Pickford, Douglas Fairbanks and D. W. Griffith, forming United Artists as an independent company to distribute their films.

The introduction of sound to the cinema, however, brought an end to Chaplin's greatness. His style of performance, derived from the circus clown and from mime, no longer seemed to work its magic. He avoided using the new technology for his films *City Lights* and *Modern Times* but embraced it in his 1936 movie, *The Great Dictator*. Though Chaplin continued to make the occasional film, and also wrote two books, his glory days were over. His leftist politics brought him in for a good deal of criticism (as did an affair with a young woman) and investigation by the Un-American Affairs Commission. As a result, Chaplin left the U.S. in 1952 and, having been refused re-entry, made his home in Switzerland. In 1972 he returned to the United States to receive several tributes, among them a special Academy Award for his contributions to the film industry. Three years later he was knighted. Chaplin died on December 25th 1977. Among his obituaries was a quote from the actor in 1960: “I remain one thing and one thing only, and that is a clown. It places me on a far higher plane than any politician.”

From: *Speak Up*. Agosto 1999 – n^o 147 (adapted)

4 - The statements in italics⁵³ below were extracted or adapted from the text. They are all correct. Choose the proposition(s) in which the statement in letter a) is correctly explained or interpreted in letter b), according to the text.

01. a) Chaplin avoided using the new technology for some of his films but embraced it in his *The Great Dictator*.
 b) Chaplin decided to introduce sound to many films but didn't accept to use it in *The Great Dictator*.
02. a) The introduction of sound to the cinema brought an end to Chaplin's greatness.

⁵³ In this research, they are all options 'a'.

- b) When silent films disappeared fame deserted Chaplin.
04. a) Chaplin's glory days were over.
b) Chaplin's fame was gone.
08. a) "It places me on a far higher plane than any politician."
b) The artist compares himself to a politician, and as a clown he feels less important.
16. a) Chaplin left the U.S. and, having been refused re-entry, made his home in Switzerland.
b) Chaplin decided to live in Switzerland because the American people finally accepted his bad manners.
32. a) Chaplin's leftist politics brought him in for a good deal of criticism.
b) Chaplin's political ideals provoked a lot of criticism against him.

5 - Select the proposition(s) which contains (contain) correct references to the following words, underlined in the text

01. which (paragraph 1): the films
02. their (paragraph 1): Mary Pickford, Douglas Fairbanks, D.W. Griffith, and Chaplin
04. its (paragraph 2): the circus clown
08. it (paragraph 2): the new technology
16. them (paragraph 2): several tributes
32. the actor (paragraph 2): Chaplin
64. that (paragraph 2): the actor

6 - Identify the correct proposition(s) according to the text

01. As Chaplin's reputation increased, so did his salary and power.
02. Chaplin could ask for a large amount of money for his movies after becoming famous.
04. For many of the films he saw, Chaplin composed the music.
08. Chaplin's style of performance was taken from the circus clown and mime.
16. After sound was introduced to the cinema, Chaplin's performance did not work its magic anymore.
32. Chaplin tried to re-enter the United States, but was not allowed. So he established himself in Europe.

TEXT 3

REWARDS FOR TALENTS



Awards and medals are usually given throughout the world to outstanding people in several areas of knowledge. One of the most famous awards is the Nobel Prize. There are other well-known premiums in the United States.

PULITZER PRIZES – they were endowed by Joseph Pulitzer (1847-1911), publisher of the New York World, in a bequest to Columbia University. They are awarded annually since 1917 for work done during the preceding year. All prizes are \$3,000 in each category (Journalism, Literature and Music), except Meritorious Public Service for which a gold medal is given.

OSCAR – a gold-plated statuette awarded by the American Academy of Motion Picture Arts and Sciences for outstanding contributions to motion-picture industry since 1928. The first movie to get an Oscar was *Wings* and the first director was Frank Borzage with *Seventh Heaven*. There are many versions about the origin of the name “Oscar”, which has been used since 1931. The most common one is that the statuette was named after Oscar Pierce, the uncle of Margaret Herrick, a librarian of the Academy.

GRAMMY – A statuette awarded annually by the National Academy of Recording Arts and Sciences for outstanding achievement in almost 70 categories in the recording industry. The first Grammy was delivered in 1958 to Domenico Modugno for his song *Volare*. The word Grammy comes from GRAM (ophone).

From: LIBERATO, Wilson Antônio. *Compact English Book*. FTD, 1998.

7 - Select the correct proposition(s) according to the text

- 01. Among the many well-known awards given in the United States, the Nobel Prize is the most famous one.
- 02. The first Pulitzer Prizes were awarded by Joseph Pulitzer, a publisher of the New York World.
- 04. Music is one of the categories awarded by both the Pulitzer Prizes and the Grammy.
- 08. The prizes mentioned in the text were all named after outstanding people.
- 16. The name Oscar was probably a tribute to Margaret Herrick’s uncle.

8 - In which paragraphs can you find the following information? Select the correct proposition(s) according to the text.

- 01. The probable origin of the name of a premium given to important contributions to the film industry: paragraph 3
- 02. The approximate amount of categories that receive a statuette in the world of the recording industry: paragraph 4
- 04. The name of a country where famous rewards are delivered: paragraph 1

08. The name of a prize that is awarded monthly since the beginning of the century:
paragraph 2
16. The year in which the name “Oscar” was first used to name a gold-plated statuette:
paragraph 3
32. How long the person who endowed the Pulitzer Prizes lived: paragraph 2
64. The price of the gold medal that is delivered as a Pulitzer Prize: paragraph 2

9 - Which of the following questions can be answered according to the information contained in the text?

01. How much do Americans spend on awards and medals given to famous people around the world every month?
02. What is the name of the artist who received a Pulitzer Prize last year?
04. What was the first song to receive a Grammy?
08. What do people win a Pulitzer Prize for?
16. Who won an Oscar for Best Director this year?
32. How many premiums are mentioned in the text?

TEXT 4

MOVIES AND PHOTOGRAPHS

If we want to learn about other societies, it is not always necessary to travel. We can discover what happens in other parts of the world by watching movies. It is difficult to imagine an easier method of learning about other countries. Nowadays movies not only tell stories or record important historical happenings. They also record for us the actions and habits of ordinary people. Much of our present knowledge of living forms and of objects in distant space, too, is obtained from movies and photographs.

From: TOTIS, Verônica. *Língua Inglesa: Leitura*. Cortez, 1991.

10 - The following is the last paragraph of the text. Select the proposition(s) that presents (present) the correct punctuation

01. Camera eyes are generally more accurate, than the eyes of men and women when a man looks at the world. He sees only what he chooses to see. He often finds it more convenient not to notice certain things. But, a camera represents every object completely and truthfully. Without this instrument many scientific discoveries. Would be impossible and we would be less sure of many historical facts.
02. Camera eyes are, generally, more accurate than the eyes of men and women. When a man looks at the world, he sees only what he chooses to see. He often finds it more convenient not to notice certain things. But a camera represents every object completely and truthfully without this instrument. Many scientific discoveries. Would be impossible and we would be less sure of many historical facts?

04. Camera eyes are generally more accurate than the eyes of men and women. When a man looks at the world, he sees only what he chooses to see? He often finds it more convenient not to notice certain things; but a camera represents every object completely and truthfully. Without, this instrument many scientific discoveries would be impossible! And we would be less sure of many historical facts.
08. Camera eyes are generally more accurate than the eyes of men and women. When a man looks at the world, he sees only what he chooses to see. He often finds it more convenient not to notice certain things. But a camera represents every object completely and truthfully. Without this instrument, many scientific discoveries would be impossible and we would be less sure of many historical facts.
16. Camera eyes are generally more accurate than the eyes of men and women. When a man looks at the world he sees only. What he chooses to see? He often finds it more convenient not to notice certain things. But a camera represents every object completely and truthfully. Without this instrument, many scientific discoveries would be impossible and we would be less sure of many historical facts!

11 - Select the correct proposition(s) to complete the following sentence: the text makes reference to...

01. Travels around the Americas.
 02. The contribution of movies and photographs to our knowledge of the world.
 04. The fact that movies and photographs can help us learn.
 08. The stories of famous people.
 16. The habits of rich people.
 32. Historical American events.
 64. An easy way to learn about other countries

12 - Which proposition(s) shows(show) the main idea of all texts, according to their sequence?

01. Text 1 – Chaplin’s beginnings and how he achieved success.
 Text 2 – Chaplin’s glory, how he lost his fame and what happened in his life until he died.
 Text 3 – Awards and medals that people receive all over the world.
 Text 4 – The importance of movies and photographs.
02. Text 1 – Chaplin’s life.
 Text 2 – The decline of silent films and Chaplin’s death.
 Text 3 – The Nobel Prize – one of the most important awards.
 Text 4 – The importance of photographs in representing knowledge.

04. Text 1 – The positive responses of cinema audiences to Chaplin’s new character.
Text 2 – The tributes received by Chaplin close to the end of his life.
Text 3 – People’s opinion about the different rewards for talents.
Text 4 – The facility of learning about other countries.
08. Text 1 – An account of Chaplin’s career and some other biographical notes about him.
Text 2 – Chaplin’s fame and decline and what happened to him up to his death.
Text 3 – Premiums given to people in different fields of activity.
Text 4 – Movies and photographs in our lives.
16. Text 1 – A description of Chaplin’s most important character.
Text 2 – Chaplin’s death.
Text 3 – The origin of some of the very well-known statuettes awarded every year.
Text 4 – The autonomy man has in choosing what he wants to see

APPENDIX 12

**Correspondence of the test tasks as presented within UNICAMP examinations
compared to the tasks required for university studies**

Analysis of the correspondence between the characteristics of TLU situation tasks and test tasks of UNICAMP examination

Characteristics of the Test tasks	Correspondence
1. Characteristics of the setting	Always low
2. Characteristics of the test rubrics	Always low
3. Characteristics of the input	Input may refer to both the texts and the tasks presented to the test taker
<i>3.1 Format</i>	
Form	Language and non-language (pictures). High
Language	Tasks are in the native language, Portuguese. Tasks = high Texts are in the target language, English. Texts = high
Length	Some are short and some are long, but not as long as academic texts. Medium
Type (text and task)	Text type: Newspaper articles, interview excerpts, poems, magazine articles, comic strips, excerpts of narratives (story books, novels, fairy tale books) and journal articles Low
	Task type: Items in the form of questions Low

To continue...

Table XX: analysis of the correspondence between the characteristics of TLU situation tasks and test tasks of UNICAMP examination (cont.)

	Characteristics of the Test tasks	Correspondence
<i>3.2 Language of input</i>	Input is to consider both the texts and the tasks presented to the test taker	
a) Language characteristics		
▪ Organizational characteristics: Grammatical	General and topic-related vocabulary Simple and complex syntax	High
▪ Organizational characteristics: Textual	Cohesion: Cohesion in some texts presented in the test is different from the cohesion in academic expository texts. Examples: the stories, comic strips, advertisement, and poems Rhetorical organization: - narratives - description, cause/consequence - problem/solution	Medium to low Low High
▪ Pragmatic characteristics: Functional	Ideational Manipulative Imaginative (poems)	High Medium Low
▪ Pragmatic characteristics: Sociolinguistic	Register – many texts with language more informal than academic texts, such as comic strips, poems, advertisement, and stories Cultural reference – texts with cultural reference. Figurative language – some text, especially the comic strips, poems, proverbs.	Low Medium Low
b) Topical characteristics	Topics: personal, cultural, curiosities, variety, scientific	Low

To continue...

Table XX: analysis of the correspondence between the characteristics of TLU situation tasks and test tasks of UNICAMP examination (cont.)

	Characteristics of the Test tasks	Correspondence
4. Characteristics of the expected response		
<i>4.1 Format</i>		
Form	Language	High
Language	Native language = Portuguese	High
Type of response	Limited production	Medium
5. Relationship between input and response.		
<i>5.2 Scope of relationship</i>	Mostly narrow scope, since most questions can be answered based on specific details or limited part of the text. Some broad scope, such as explain the title.	High
<i>5.3 Directness of relationship</i>	Direct and indirect – some response can be given using primarily the information provided by the input, but some cannot.	High

Based on the following information about each text:

Text 1

A short story involving a dialog, written by Philip Ridley, published the 1996.

- Language of the input
 - o Vocabulary: general and specific related to life in a small city
 - o Syntax: simple and complex structures
 - o Functional: mainly imaginative
 - o Sociolinguistic: low degree of formality
 - o Topic: the construction of a building and the changes it caused

Text 2

An expository text, titled *The Soil-eaters*, presenting some factual information and a dialog with the reader, published on the Internet by Nature News Service in 1996

- Language of the input
 - o Vocabulary: general and specific related to eating
 - o Syntax: simple and complex structures
 - o Functional: ideational and manipulative
 - o Sociolinguistic: middle degree of formality
 - o Topic: eating and geophagy

Text 3

A small expository text, no titled, presenting findings of a scientific research on the elderly's health, published by New Scientist on September, 1991.

- Language of the input
 - o Vocabulary: general and specific related to health
 - o Syntax: simple and mainly ideational complex structures
 - o Functional:
 - o Sociolinguistic: middle degree of formality
 - o Topic: research findings on the elderly's health

Text 4

02 Letters to the editor??, one titled *Murphy was a Perfectionist*, presenting an appraisal of a previously published article on Murphy and Murphy's law, as well as some description of factual information about the law, and the other, no title, presenting an example of Murphy's law by one reader of the magazine, both published on the August 1997 issue of the Scientific American.

- Language of the input
 - o Vocabulary: general
 - o Syntax: simple and complex structures
 - o Functional: mainly ideational
 - o Sociolinguistic: middle degree of formality
 - o Topic: reference to a previous article on Murphy's law

Text 5

An expository text, titled Caesar's Ghost: the real reason why things never change, presenting factual information about the development of railroads, published on UTNE Reader, July-August 97

- Language of the input
 - o Vocabulary: general and specific related to railroads
 - o Syntax: simple and complex structures
 - o Functional: ideational
 - o Sociolinguistic: middle degree of formality and cultural reference
 - o Topic: the development of railroads and its influence on the distance between the rails

APPENDIX 13

Segue-se um trecho de uma história retirada de *The Victorian Fairy-Tale Book*. Leia-o e responda às questões 8 e 9.

A great fear came over the poor boy. Lonely as his life had been, he had never known what it was to be absolutely alone. A kind of despair seized him – no violent anger or terror, but a sort of patient desolation.

“What in the world am I to do?” thought he, and sat down in the middle of the floor, half inclined to believe that it would be better to give up entirely, lay himself down, and die.

This feeling, however, did not last long, for he was young and strong, and, I said before, by nature a very courageous boy. There came into his head, somehow or other, a proverb that his nurse had taught him – the people of Nomansland were very fond of proverbs –

For every evil under the sun
There is a remedy, or there's none;
If there is one, try to find it-
If there isn't, never mind it.

“I wonder – is there a remedy now, and could I find it?” cried the Prince, jumping up and looking out of the window.

8. Em que situação se encontrava o protagonista da história e o que ele pensava em fazer inicialmente?

9. Explique como ele chega a mudar de idéia.

APPENDIX 14

Leia o trecho seguinte, do livro *The Love You Make. An Insider's Story of The Beatles*, de P. Brown e S. Gaines (trecho em que são mencionados John Lennon, sua mãe Júlia, sua tia Mimi e seu pai Fred) e responda à questão 2.

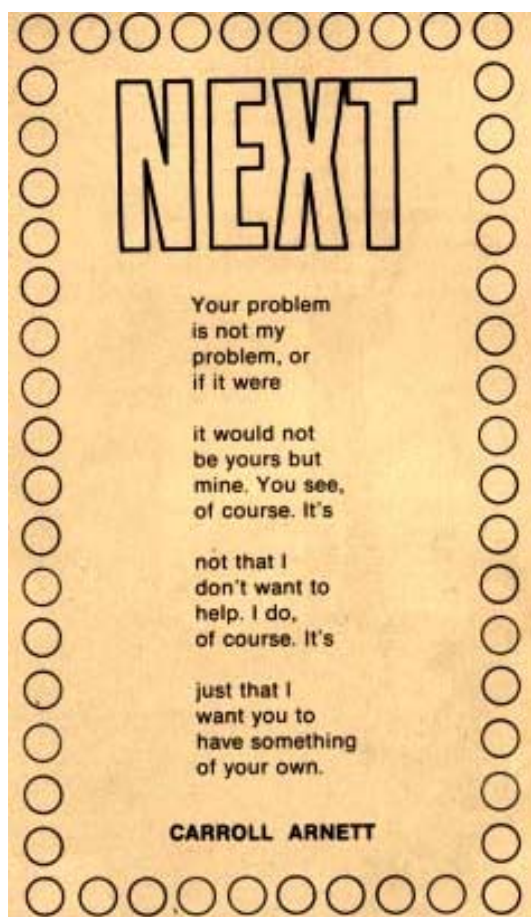
(...) But by that summer it had become clear that John wasn't interested in his education, or in art, or in his future at all. John's only interest was the American craze called "rock and roll", a derivative form of black rhythm and blues with a prominent drum beat. (...)

John wanted a guitar more than he had wanted anything before in his life. Surprisingly, it wasn't Julia who broke down and bought it for him, it was Mimi who marched him to a music shop in Whitechapel and bought him his first guitar for £17. A small, Spanish model with cheap wire strings, he played it continuously until his fingers bled. Julia taught him some banjo chords she had learned from Fred, and he started with those. He sat on the bed all day, and when Mimi tried to shoo him into the sunlight, he'd go out to the support and lean up against the brick wall practicing his guitar for so long that Mimi thought he'd rub part of the brick away with his behind. She watched him waste hour after hour, day after day with the damned thing and regretted having bought it for him. "The guitar's all very well, John," she warned him, "but you'll never make a living out of it."

2. Qual a previsão feita por Mimi a respeito do futuro de John Lennon?

APPENDIX 15

Leia o poema abaixo e responda à questão 16.



Poema originalmente publicado em *Not only that* (The Elizabeth Press, 1967) e reproduzido em M.L.Greene (ed.) *Another Eye*. Illinois, Scott, Foresman and Company, 1971, p. 121.

16. Como o poema de Carroll Arnett justifica que *Your problem is not my problem*?

APPENDIX 16

O que se segue são os parágrafos iniciais de “Ghosts”, um conto de Paul Auster publicado em *The New York Trilogy*, em 1990, pela Penguin Books Inc. Leia-os e responda à pergunta 14.

FIRST of all there is Blue. Later there is White, and then there is Black, and before the beginning there is Brown. Brown broke him in, Brown taught him the ropes, and when Brown grew old, Blue took over. That is how it begins. The place is New York, the time is the present, and neither one will ever change. Blue goes to his office every day and sits at his desk, waiting for something to happen. For a long time nothing does, and then a man named White walks through the door, and that is how it begins.

The case seems simple enough. White wants Blue to follow a man named Black and to keep an eye on him for as long as necessary. While working for Brown, Blue did many tail jobs, and this one seems no different, perhaps even easier than most.

Blue needs the work, and he listens to White and doesn't ask many questions. He assumes it's a marriage case and that White is a jealous husband. White doesn't elaborate. He wants a weekly report, he says, sent to such and such a postbox number, typed out in duplicate on pages so long and so wide. A check will be sent every week to Blue in the mail. White then tells Blue where Black lives, what he looks like, and so on. When Blue asks White how long he thinks the case will last, White says he doesn't know. Just keep sending the reports, he says, until further notice.

14. Quais são os personagens que aparecem nesse trecho? Como esses personagens se interrelacionam?

APPENDIX 17

Responda a todas as perguntas EM PORTUGUÊS.

Leia o trecho abaixo e responda às questões **01**, **02** e **03**.

Day by day the Point got taller and taller. And day by day the shadow got longer and longer.

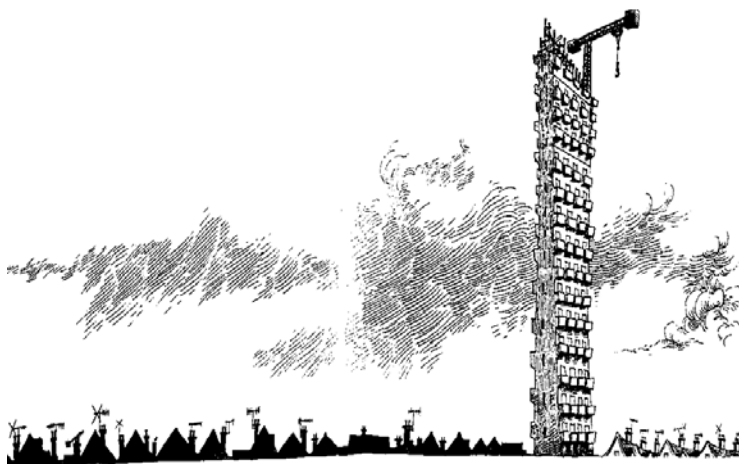
All around flowers died, grass turned brown and rooms became dark and cold. Old people had to turn on heaters, even in the middle of summer.

‘It’s just so ugly,’ said Doll to Harold as they ate dinner one night. ‘Once I used to look out of the window and see trees and flowers, hear singing birds. Now all I see is that ugly grey thing. There’re no flowers, no trees, no light, no grass, no birds, nothing.’

‘Oh, it’s not that bad,’ said Harold.

‘Don’t give me that,’ snapped Doll. ‘You don’t have to watch it. Day in and day out. Watch it getting bigger and bigger and bigger.’

Rosie sat at the table and ate her dinner. She thought her mum was being stupid, although she didn’t say so. Instead, she just filled her mouth with a forkful of mashed potato and stared at her plate.



Later, though, while Doll was washing up, Rosie couldn’t help saying, ‘I don’t think it’s ugly.’ ‘Well, you’re as foolish as your father, then.’ ‘I just think it’s . . . it’s a gigantic finger pointing up to the sky. Or a tall flower. Or a wonderful steeple –’

‘Listen, young lady,’ interrupted Doll. ‘It’s not a finger and it’s not a flower and it’s not a steeple. It’s just a shadow. Nothing else. It’s just a point of shadow.’

And that was how the Point became known as Shadow Point.

(*Philip Ridley. Mercedes Ice. London, Puffin Books. 1996, pp. 18-19*)

01. Quem é quem nessa história?

02. A que se refere “*Shadow Point*”? Por que recebeu esse nome?

03. O texto menciona mudanças. Que mudanças são essas?

As questões 04, 05 e 06 dizem respeito ao texto abaixo.

nature science update

[Update] [Next Article]

The soil-eaters

By Ehsan Masood

nature

It's lunchtime somewhere in rural tropical Africa. You're hungry, but the nearest restaurant is too far to walk. There's no Italian, Chinese, Indian or fast food and the telephone pizza delivery company is a little reluctant to send its dispatch rider beyond the city walls. Moreover, you're on a tight budget. What are you to do? The answer, quite literally, may lie in the soil directly beneath your feet.

According to two researchers from the University of Wales at Aberystwyth, UK, the tradition of soil consumption is still very much alive in the African tropics, India, Jamaica and it has also been reported in Saudi Arabia.

Despite the advent of modern religions and the end of the slave trade, soil eating is not uncommon, though mostly confined to the poorer sections of society.

The reasons for soil consumption are many and often misunderstood, say the researchers Peter Abrahams and Julia Parsons. But geophagists – as soil-eaters are known – on the whole are regarded as quite 'normal' to most but outsiders.

"Despite the widespread distribution of geophagy, both today and in the past, it is largely unknown, under-reported, misunderstood or ignored by most people in the developed world", say Abrahams and Parsons. [This is why] "the adjectives 'eccentric', 'perverted', 'odd', and 'bizarre' have all been applied to geophagy".[...]

(Nature News Service, 1996)

04. O primeiro parágrafo se dirige a um público-leitor específico. Que público é esse? Justifique sua resposta.

05. Qual é a explicação de Abrahams e Parsons para o uso de adjetivos como "eccentric", "perverted", "odd" e "bizarre" para caracterizar a geofagia?

06. Dê um significado para a palavra "but" no trecho "...on the whole [soil eaters] are regarded as quite 'normal' to most but outsiders".

Leia o texto abaixo e responda à questão **07**.

A SIDELIGHT on urban violence in the US could also be showing up a similar situation in some parts of the UK. A doctor in Arkansas has pointed out that the rise of street gangs is affecting preventive medicine for elderly people. He mentioned two patients of his, both in their early 60s, one with hypertension and the other with diabetes. Both took regular walks of a mile or two several times a week, but they have become too frightened of street gangs to go out.

Their walks ceased several months ago. Consequently both had gained about 10 pounds in weight, not a good thing for either condition. So street gangs, apart from the obvious damage they can cause, might also be worsening cardiovascular disease and diabetes in the elderly. I do not know whether anyone has noticed gains in weight for the same reason among elderly patients in some parts of London, for example.

Bill Tidy

(*New Scientist* 28 September 1991)

07. De que maneira a violência urbana pode estar afetando a saúde de pessoas idosas?

Leia os dois textos abaixo, da seção *Letters*, e responda às questões **08, 09, 10 e 11**.

LETTERS

MURPHY WAS A PERFECTIONIST

As the son of the man whose name is attached to “Murphy’s law,” I want to thank you for accurately and respectfully identifying the origin of this “law” in your recent article [“The Science of Murphy’s Law,” by Robert A.J. Matthews, April]. My father was an avid reader of *Scientific American*, and I can assure you that were he still alive, he would have written to you himself, thanking you for a more serious discussion of Murphy’s Law than the descriptions on the posters and calendars that treat it so lightly.

Yet as interesting as the article is, I suggest that the author may have missed the point of Murphy’s Law. Matthews describes the law in terms of the probability of failure. I would suggest, however, that Murphy’s law actually refers to the CERTAINTY of failure. It is a call for determining the likely causes of failure in advance and acting to prevent a problem before it occurs. In the example of flipping toast, my father would not have stood by and watched the slice fall onto its buttered side. Instead he would have figured out a way to prevent the fall or at least ensure that the toast would fall butter-side up.

Murphy and his fellows engineers spent years testing new designs of devices related to aircraft pilot safety or crash survival when there was no room for failure (for example, they worked on supersonic jets and Apollo landing craft). They were not content to rely on probabilities for their successes. Because they knew that things left to chance would definitely fail, they went to painstaking efforts to ensure success.

EDWARD A. MURPHY III, Sausalito, California

After receiving more than 362 intact issues of *Scientific American*, I received the April issue – with the article on Murphy’s Law – that was not only assembled incorrectly by the printer but also damaged by the U.S. Post Office during delivery. My teenage daughter is taking this magazine into her science class to talk about Murphy’s Law. The condition of this issue is an excellent example for her presentation.

BRAD WHITNEY, Anaheim, California

(*Scientific American*, August 1997)

08. O que deu origem a esses dois textos?

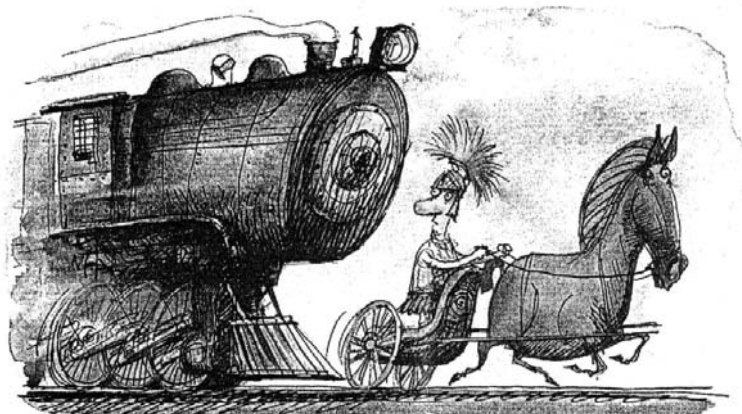
09. O primeiro texto destaca dois pontos positivos e faz uma ressalva. Transcreva o quadro abaixo para o seu caderno de respostas, preenchendo-o com as informações necessárias:

Pontos positivos	1.
	2.
Ressalva	

10. O segundo texto afirma: “The condition of this issue is an excellent example for her presentation”. Explique por quê.

11. Explique por que Murphy pode ser considerado um perfeccionista.

Leia o texto abaixo e responda à questão 12.



Caesar's Ghost

The real reason why things never change

The U.S. standard railroad gauge – the distance between the rails – is 4 feet, 8.5 inches. Why that exceedingly odd number? Because that's the way they built them in England, and the U.S. railroads were built by English expatriates. Why did the English people build them like that? Because the first rail lines were built by the same people who built the prairailroad tramways, and that's the gauge they used. Why? Because the people who built the tramways used the same jigs and tools for building wagons, which used that wheel spacing. OK! Why did the wagons use that odd wheel spacing? Well, if they tried to use any other spacing their wagons would break on some of the old long-distance roads, because that's the spacing of the old wheel ruts.

RICHARD THOMSON

(**UTNE READER**, *July-August 97*, p. 32)


So who built the old rutted roads? The first long-distance roads in Europe were built by Imperial Rome for the benefit of their legions and have been used ever since. The initial ruts, which everyone else had to match for fear of destroying their wagons, were first made by Roman war chariots, which, because they were made for or by Imperial Rome, were all alike in the matter of wheel spacing. So, the U.S. standard railroad gauge of 4 feet, 8.5 inches derives from the original specifications for an Imperial Roman army war chariot. Specs and bureaucracies live forever.

From Kyoto Journal (#33). Subscriptions: \$40 for 4 issues from 31 Baud St., York, NY 10012.

12. Explique o título desse texto.

APPENDIX 18



Leia o texto abaixo, propaganda de uma companhia de energia elétrica nos Estados Unidos, e responda à questão 10.



WHEN 93-YEAR OLD WARD NEWTON CALLED SAYING HE COULDN'T SLEEP WITH HIS POWER OUT, I KNEW I COULDN'T EITHER.

"IT WAS ABOUT 2 A.M. WHEN OUR CUSTOMER SERVICE REPRESENTATIVE, MYRA WATSON, GOT THE CALL. IT WASN'T AN EMERGENCY, BUT I GUESS HE SOUNDED REALLY WORRIED SO SHE ASKED ME TO GO OUT AND HAVE A LOOK. I GOT THERE IN ABOUT 30 MINUTES. IT SEEMS A DELIVERY TRUCK HAD BACKED INTO MR. NEWTON'S HOUSE, KNOCKIN' LOOSE A CONNECTION. IT TOOK LESS THAN HALF AN HOUR TO FIX. HE WAS STILL A LITTLE NERVOUS, THOUGH, SO I HELPED HIM RESET HIS CLOCKS, THEN SAT WITH HIM A FEW MINUTES 'TIL HE FELT BETTER. TURNED OUT HE'D KNOWN MY FATHER. WELL, AS I'M LEAVING HE SAYS NOW THAT THE POWER IS ON, HE CAN TURN OUT THE LIGHTS. I KNEW WHAT HE MEANT."

AT ENTERGY, WE'VE GOT LOTS OF PEOPLE LIKE MYRA WATSON AND SERVICEMAN DAVID BELL. PEOPLE WHO'LL DO JUST ABOUT ANYTHING TO TAKE CARE OF THEIR CUSTOMERS. IF YOU'VE GOT A QUESTION, CALL US AT 1-800-ENTERGY. AND DISCOVER THE POWER OF PEOPLE.



WWW.ENERGY.COM
© 1997 ENTERGY CORPORATION PAID FOR BY ENTERGY SHAREHOLDERS.

10. Qual era o problema do Sr. Newton?

APPENDIX 19

Taxonomies of Reading Skills

Munby's (1978) taxonomy of reading skills or sub-skills:

- Recognizing the script of the language
- Deducing the meaning and use of unfamiliar lexical items
- Understanding explicitly stated information
- Understanding information when not explicitly stated
- Understanding conceptual meaning
- Understanding the communicative value of sentences
- Understanding relations within the sentence
- Understanding relations between parts of a text through lexical cohesion devices
- Understanding cohesion between parts of a text through grammatical cohesion devices
- Interpreting text by going outside it
- Recognizing indicators in discourse
- Identifying the main point or important information in discourse
- Distinguishing the main idea from supporting details
- Extracting salient details to summarize (the text, an idea)
- Extracting relevant points from a text selectively
- Using basic reference skills
- Skimming
- Scanning to locate specifically required information
- Transcoding information to diagrammatic display

Rosenshine's (1980) findings of the common comprehension skills:

1- locating details

- Recognition
- Paraphrase
- Matching

2- simple inferential skills

- Understanding words in context
- Recognizing the sequence of events
- Recognizing cause and effect relationships
- Comparison and contrasting

3- complex inferential skills

- Recognizing the main idea/title/topic
- Drawing conclusions
- Predicting outcomes

List of comprehension subskills or underlying skills proposed for reading literature books published by Ginn and Company grade series (Rosenshine, 1980):

- Matching character with traits, actions, and speech
- Classifying questions about a selection
- Categorizing story elements into problem, climax, and solution
- Listing characters to match given dialogue or actions
- Giving setting and time
- Stating the moral
- Stating point of view from which the story is told
- Recounting character traits, qualities
- Giving an account of similarities or differences in the content or plot of selections
- Explaining the suitability of titles and heading
- Making inferences about what would happen if circumstances were different
- Evaluating ideas in a selection
- Matching events to time.

List of skills for critical reading (Alderson, 2000):

- Evaluate deductive inferences
- Evaluate inductive inferences
- Evaluate the soundness of generalization
- Recognize hidden assumptions
- Identify bias in statements
- Recognize author's motives
- Evaluate strength of arguments

List of abilities of a good reader (Grabe, 1999):

- Fluent and automatic word recognition skills, ability to recognize word parts (affixes, word stems, common letter combinations)
- A large recognition vocabulary
- Ability to recognize common word combinations (collocations)
- A reasonably rapid reading rate
- Knowledge of how the world works (and of the L2 culture)
- Ability to recognize anaphoric linkages and lexical linkages
- Ability to recognize syntactic structures and parts of speech information automatically
- Ability to recognize text organization and text-structure signaling
- Ability to use reading strategies in combination as strategic readers
- Ability to concentrate on reading extended texts
- Ability to use reading to learn new information
- Ability to determine main ideas of a text
- Ability to extract and use information, to synthesize the information, to infer information
- Ability to read critically and evaluate text information.

List of strategies and skills for academic purposes used for the development of the IELTS (Clapham, 1996):

- Identifying structure, content, sequence of events and procedures
- Following instructions
- Finding main ideas which the writer has attempted to make salient
- Identifying the underlying theme or concept
- Identifying ideas in the text, and the relationships between them, e.g. probability, solution, cause, effect
- Identifying, distinguishing and comparing facts, evidence, opinions, implications, definitions and hypotheses
- Drawing logical inferences
- Evaluating and challenging evidence
- Formulating an hypothesis from underlying theme, concept and evidence
- Reaching a conclusion by relating supporting evidence to the main idea

APPENDIX 20

UFSC SPECIFICATIONS

The following specifications are taken from the candidate's manual, and have been summarized and translated by the researcher:

- Objective: to focus on the use of language.
- Concept of reading: global comprehension of main points, and local comprehension for details
- Vocabulary: candidate must demonstrate knowledge of basic vocabulary;
Grammar: assessed indirectly, as an accessory to comprehension
- Type of texts: authentic and simplified or simple?
- Tasks to elicit reading performance in order to gather the relevant information in terms of reading skills for interpretation of reading competence/ability:
 - a) identify the types of texts XX;
 - b) use strategies of scanning and skimming
 - c) recognize main topics and secondary topics/details
 - d) identify ideas and the existing relations among them
 - e) locate key words
 - f) use visual information as aid to textual comprehension
 - g) recognize words and expressions with similar meanings
 - h) identify contextual reference
 - i) read carefully seeking logical conclusions;
 - j) relate information, seeking the intertextuality
 - k) demonstrate adequate knowledge of grammatical structure which allows for the comprehension of the texts in the tests
- Text subject matter: various topics.
- Text types: various text types.

APPENDIX 21 UNICAMP SPECIFICATIONS

The following specifications are taken from the candidate's manual. They have been summarized and translated by the researcher and are presented in an organization to serve the purpose of the research:

- objective: to assess reading competence in a foreign language, since this ability is essential for a university student to carry out their studies.
- concept of reading: reading is not a passive decoding of meaning, but an active task of negotiating meaning based on global comprehension, resulting in a new text by the reader (with limits, since not any reading is allowed).
- focus of the questions: information present in the text, and information underlying its structure. Grammar knowledge is not explicitly tested, i.e., the candidates "will not find questions on discrete grammatical points, as, for example, verb conjugation, preposition use, etc." (Candidate's Manual).
- tasks to elicit reading performance in order to gather the relevant information in terms of reading skills for analysis of reading competence: a) identify and extract information the way it appears in the text; b) put the information in order in a way to distinguish what is relevant and the irrelevant; c) identify the existing relations between two or more elements within the text; d) locate segments of the text to justify an answer or transcribe segments to account for a certain aspect of the text; e) reconstruct the controlling idea articulating some pieces of information; f) identify segments of the text conveying value judgement about information present in the text; g) recognize some elements part of the discursive nature of the text, such as the identification of the author, the audience, and point of view; h) show the ability to guess the meaning of words and expressions; i) determine the consequences of the choice and use of some words and expressions in their contexts; j) identify relations and contradictions between and among texts; k) identify discourse markers such as *it is important to...*, *finally*, *however*, and *this* and *that*; and l) the identification of text writer and audience, the context, the objective, the media, titles,

subtitles, letter type, and extralinguistic features, such as pictures, photos, graphs and illustrations.

- text subject matter: various topics, chosen within the candidates' background knowledge, not limited to some specific domain.
- text types: various text types, written with standard English, providing the reader with different types of discursive experience and degrees of reading difficulty.