JYVÄSKYLÄ STUDIES IN HUMANITIES

159

Joanna Anckar

Assessing Foreign Language Listening Comprehension by Means of the Multiple-Choice Format: Processes and Products

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN HUMANITIES 159

Joanna Anckar

Assessing Foreign Language Listening Comprehension by Means of the Multiple-Choice Format: Processes and Products

Esitetään Jyväskylän yliopiston humanistisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Villa Ranan Paulaharjun salissa
syyskuun 17. päivänä 2011 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Humanities of the University of Jyväskylä,
in building Villa Rana, Paulaharju hall, on September 17, 2011 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2011

Assessing Foreign Language Listening Comprehension by Means of the Multiple-Choice Format: Processes and Products

JYVÄSKYLÄ STUDIES IN HUMANITIES 159

Joanna Anckar

Assessing Foreign Language Listening Comprehension by Means of the Multiple-Choice Format: Processes and Products

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2011

Editors
Mirja Tarnanen
Center of Applied Language Studies, University of Jyväskylä
Pekka Olsbo, Ville Korkiakangas
Publishing Unit, University Library of Jyväskylä

URN:ISBN:978-951-39-4410-0
ISBN 978-951-39-4410-0 (PDF)

ISBN 978-951-39-4409-4 (nid.)
ISSN 1459-4323 (nid.), 1459-4331 (PDF)

Jyväskylä University Printing House, Jyväskylä 2011

ABSTRACT

Anckar, Joanna
Assessing Foreign Language Listening Comprehension by Means of the Multiple-Choice Format: Processes and Products
Jyväskylä: University of Jyväskylä, 2011, 307 p.
(Jyväskylä Studies in Humanities
ISSN 1459-4323 (nid.), 1459-4331 (PDF); 159)
ISBN 978-951-39-4409-4 (nid.)
ISBN 978-951-39-4410-0 (PDF)
Finnish summary.
Diss.

In this study, the processes and strategies behind test-takers' performance on MC items are investigated. The starting points for the study are: the cognitive nature of the listening process, the nature of the MC format, item validation and the introspection method. The analyzed 17 MC items assessing L2 French listening comprehension come from a high-stakes context: the Finnish Matriculation Examination (spring 2002). The main research tool is "short written introspection": 218 test-takers on the upper secondary school level are asked to justify their option selection for each item. An analysis of the contents of the items and a Rasch analysis are conducted to justify the further analysis of the items. The introspective responses provide a covering picture of the processes and strategies activated in a MC test situation. They range from evidence of comprehension on different levels of the spoken text and of the use of strategies to affective reactions to the task at hand. Detailed information is obtained for example on the nature of guessing and elimination and their use as a function of the characteristics of individual test-takers and items. The study points at flaws in the items that sometimes seem difficult to foresee: excessive textual information load, opaque questions or options and implausible or not clearly false options prove to affect the test-takers' processing and their choice of strategies. Items with flaws often fail to measure the targeted construct and to discriminate between test-takers with and without the targeted ability. These items represent threats against the reliability of item scores and against the validity of their use.

Keywords: multiple-choice format, verbal protocol, L2 listening comprehension, item validation, language assessment

Author's address
Joanna Anckar
Soveltavan kielentutkimuksen keskus
University of Jyväskylä
janckar@abo.fi

Supervisors
Dr Mirja Tarnanen
Center of Applied Language Studies
(Soveltavan kielentutkimuksen keskus)
University of Jyväskylä

Reviewers
Dr Tineke Brunfaut
Dr Pekka Rantanen
Dr Gary Buck

Opponents
Dr Tineke Brunfaut

FOREWORD

As the writing of my PhD thesis has been a very long project, there are numerous people to whom I want to express my gratitude. To make a long story short:

Thank you my supervisor professor Mirja Tarnanen. Not only are you a smart woman and a professional and inspiring coach, but a wonderful human being.Thank you my statistical advisor Reeta Neittaanmäki. You have shown an admiring amount of patience with a statistical analphabet like me! Thank you professor Urpo Nikanne who helped me with kind advice and support especially in the beginning of my long project!

Thank you to the reviewers of my thesis dr Pekka Rantanen, dr Gary Buck and dr Tineke Brunfaut. You have pointed out many important issues and suggested necessary changes for the final version. A second thank you to dr Tineke Brunfaut, who also accepted to come to Finland to be my opponent.

Thank you dr Ari Huhta for many kind advices along the road, and for giving comments to one version of the thesis. Thank you professor Sauli Takala for many advices and inspiration in many various national and international assessment contexts. Thank you Päivi Sihvonen and Eevi Nivanka for letting me interview you.

Thank you Dad for being the one who led me into the exciting world of language assessment in the first place, and for reading through one version of my thesis to get at the worst language mistakes…

Many thanks to CALS (SOLKI) at the University of Jyväskylä for having given both moral and economical support to help me get through this project! Thank you all wonderful professors and heads of department: Minna-Riitta Luukka, Tarja Nikula and Riikka Alanen. Thank you to the Finnish Matriculation Examination who has given me the opportunity to use their material and data. A big thank you to the 226 anonymous pupils and their teachers in 22 schools who agreed to take part in my research by responding to the listening comprehension items. Thank you to Doctoral school Langnet for offering a wonderful platform to meet peers and experts from all over the country! Thank you EALTA and the conferences for giving me the opportunity to present my research project at different stages and from different angles.

Thank you all colleagues and fellow researchers from different departments and different universities who have given the possibility to discuss research and practice - their joys and troubles - on many different occasions! Thank you all colleagues in the French division of the Matriculation Examination for challenging and interesting work and cooperation – where theory meets practice…

Thank you my dear family and friends for giving loving support and for often helping me keep my mind OFF the thesis!

I want to dedicate this thesis to my late grandma Majlis, who did not always have many options in her life, and to my daughters Sandra, Ellen and Nadja, who will have all the options in the world!

Pargas 7.8.2011
JoannaAnckar

FIGURES

FIGURE 1 Overview of the main focus of the thesis 21
FIGURE 2 Simplified processing model...62
FIGURE 3 The potential impact of a test 74
FIGURE 4 Proportion of different grades given in the Finnish Matriculation Examination..116
FIGURE 5 Distribution of total scores in the current test 160
FIGURE 6 Person-item map showing the distribution of item difficulty and person ability on the same scale 301
FIGURE 7 Relationship between item measure and discrimination index of 17 items 163
FIGURE 8 Trace lines for the selection of the key (2c) and the distractors (2a and 2b) 164
FIGURE 9 Trace lines for the selection of the key (11a) and the distractors 11b and 11c 165
FIGURE 10 Trace lines for the selection of the key (25c) and the distractors 25a and 25b...166
FIGURE 11 Trace lines for the selection of the key or a distractor for item 25 across the individual total score groups........................166
FIGURE 12 Trace lines for the selection of the key (3c) and the distractors 3a and 3b...167
FIGURE 13 Proportion of different responses to all items by all test-takers 171
FIGURE 14 Hypothetical optional processes and strategies at stake in a MC test-taking situation 304
FIGURE 15 Number of word-bound responses and success per item 175
FIGURE 16 Number of partial comprehension and success per item 181
FIGURE 17 Number of résumé responses and success per item........................185
FIGURE 18 Number of nonsense responses and success per item.....................189
FIGURE 19 Number of cases of the ten types of guessing..................................195
FIGURE 20 Guessing continuum 202
FIGURE 21 Type of guess and proportion of success 202
FIGURE 22 Total number of guessing and success per item...............................204
FIGURE 23 Total number of elimination and success at items 1-11 and 25-30 217
FIGURE 24 Item 1: introspective responses combined with the three options 236
FIGURE 25 Item 3: introspective responses combined with the three options 238
FIGURE 26 Item 6: Introspective responses combined with the three options 239
FIGURE 27 Item 7: Introspective responses combined with the three options 240
FIGURE 28 Item 8: Introspective responses combined with the three options 241

FIGURE 29 Item 9: introspective responses combined with the three options 242
FIGURE 30 Item 11: introspective responses combined with the three options 243
FIGURE 31 Item 25: introspective responses combined with the three options 246
FIGURE 32 Item 27 introspective responses combined with the three options 248
FIGURE 33 Item 29 introspective responses combined with the three options 251
FIGURE 34 Item 30 introspective responses combined with the three options 253
FIGURE 35 The focus in the test-taking process from the text to the test-taking strategies 256

TABLES

TABLE 1 Cognitivist and social views of SLA according to Larsen-Freeman (2007) 24
TABLE 2 Models of the listening comprehension process and difficulties associated with each stage 32
TABLE 3 A framework for describing listening ability (adapted from Buck 2001:104 38
TABLE 4 Listening comprehension strategies as classified by Young (1997), Goh (2002) and O'Malley & Chamot (1990) 51
TABLE 5 Hypothetical description of a skilled listener vs. a less skilled listener..55
TABLE 6 Advantages and disadvantages of authentic/alternative and performance assessment (Takala 1998)..59
TABLE 7 A model of information processing (goal-process system) at a multiple-choice test of listening (Jamieson et al. 2000) 83
TABLE 8 Procedures for establishing quality-based evidence (Weir 2005: 233) 88
TABLE 9 Independent text, item and text & item variables and their hypothesized impact on item difficulty (From: Rupp, Garcia & Jamieson 2001) 89
TABLE 10 Three steps of construct validation (From Haladyna 2004: 18)........91
TABLE 11 Item review activities (from Haladyna 2004: 201 96
TABLE 12 Comparison of the levels of the Finnish National Certificates of Language Proficiency and the National Language Examination (for Civil servants) 113
TABLE 13 Numbers of test-takers in the Finnish Matriculation Examination for different foreign languages in spring 2011 114
TABLE 14 The target levels for the different skills for different L2 114

TABLE 15 Number and proportion of test-takers in the original and the current test 130
TABLE 16 Chi-square analysis of 17 items: results from original and current test administration 134
TABLE 17 Proportion of selection of options for each item in the original and the current test administration 302
TABLE 18 Characteristics of the verbal protocol used in the current study ...138
TABLE 19 Summary of 216 measured (non-extreme) persons 160
TABLE 20 Results of the Rasch analysis of the 17 listening comprehension test items 303
TABLE 21 The selection of options (key/distractor) for each of the 17 listening comprehension test items related to the test-takers' person measures 161
TABLE 22 Division of test-takers into three score groups 164
TABLE 23 Description of the different introspective response categories 169
TABLE 24 Agreement on the coding of item 3 by three independent coders 173
TABLE 25 Correlation between estimated person measure and the number of the different introspective responses per test-taker (N=218) 305
TABLE 26 The different types of introspective responses correlated with the item measure 305
TABLE 27 Guesses correlated with item measure 306
TABLE 28 Subtypes of elimination, their number and proportion among all cases of elimination 209
TABLE 29 Subtypes of elimination and proportion of correct responses (in descending order) 215
TABLE 30 Elimination correlated with item measure 306
TABLE 31 Number of different metacognitive responses for the different items 306
TABLE 32 Changes of option choices between the distractors and the key after the second listening 221
TABLE 33 Correlations between the number of changes of options with item measure for the 17 items 307
TABLE 34 The different types of metacognitive responses correlated with the item measure 307
TABLE 35 Problems in different items reflected in the introspective Responses 235
TABLE 36 Reasons for selecting the key or a distractor 265

CONTENTS

ABSTRACT
FOREWORD
FIGURES AND TABLES
CONTENTS

INTRODUCTION 13
Orientation 14
Research questions 17
Data & methodology 17
Organisation of the thesis 19

1. THE LISTENING PROCESS 22
1.1 The cognitive and social nature of L2 listening comprehension 23
1.2 The role of listening comprehension in L2 teaching 29
1.3 Models of the listening comprehension process 30
1.4 Cognitive factors related to the L2 listening process 35
1.5 A listening framework for the listening construct 37
1.5.1 Grammatical knowledge in listening – processing and problems 39
1.5.2 Discourse knowledge in listening – processes and problems .. 45
1.5.3 Pragmatic and sociolinguistic knowledge 47
1.6 Listening comprehension strategies 48
1.6.1 Strategies vs. processes 48
1.6.2 Cognitive vs. metacognitive strategies 49
1.6.3 Listening comprehension strategies 50
1.7 The effect of the learner's/listener's level of skill and personal characteristics on the use of strategies 55
1.8 Test-taking strategies in listening comprehension contexts 57
1.9 Listening comprehension tasks 58
1.10 The implications for the current study of the description of listening processes and strategies 61

2 THE CONSTRUCT, THE TEST FORMAT AND THE VALIDATION PROCESS 64
2.1 The test context and the construct 64
2.2 Task parameters 67
2.3 Test usefulness 68
2.3.1 Construct validity 69
2.3.2 Reliability 70
2.3.3 Authenticity and interactiveness 72
2.3.4 Impact and washback 73
2.3.5 Practicality 75
2.3.6 Relative importance of different test qualities 76

2.4 Test and item formats ... 76
2.4.1 Multiple-choice items and other formats ... 78
2.4.2 The importance of the quality of the MC item ... 81
2.4.3 Multiple-choice test-taking strategies: guessing and elimination ... 84
2.5 Central issues on the usefulness of the test and the test format ... 85
2.6 Multiple-choice item development and validation ... 86
2.6.1 Advice given to creators of multiple-choice test items ... 91
2.6.2 Reviewing the multiple-choice items ... 96
2.6.3 Quantitative item validation ... 98
2.6.4 Distractor evaluation ... 100
2.7 Validation in the context of the present research study ... 101

3 INTROSPECTION IN THEORY AND PRACTICE IN THE ANALYSIS OF A MULTIPLE-CHOICE TEST OF LISTENING COMPREHENSION 103
3.1 Theory: Why introspection? ... 103
3.1.1 Cognitive justification for introspection ... 104
3.1.2 The typology of verbal protocols ... 104
3.1.3 Limitations of the use of introspection ... 106
3.2 Practice: Previous studies using the method of introspection ... 107

4 THE BACKGROUND TO THE TESTING CONTEXT ... 111
4.1 Language assessment in Finland ... 111
4.1.1 The Finnish Matriculation Examination ... 113
4.1.2 The original testing context: the foreign language test ... 115
4.2 The test constructors' point of view: background on the test context by two item writers & test developers ... 119
4.2.1 General information ... 120
4.2.2 The procedure of constructing a test of listening comprehension ... 120
4.2.3 Other comments on the listening comprehension test ... 122
4.3 The original test procedure of the subtest of French listening comprehension ... 123
4.4 A summary of the original test context and the construct of listening comprehension ability ... 124

5 THE RESEARCH STUDY ... 127
5.1 Research questions ... 127
5.2 Data and collection ... 127
5.2.1 The test procedure for the current research ... 128
5.2.2 Remarks on the test procedure and the study ... 130

6 METHODS OF ANALYSING THE CURRENT TEST ITEMS ... 136
6.1 Ways of analyzing the test contents ... 136
6.2 Statistical information on the test ... 137
6.3 The use of introspective analysis in the current study ... 138

ANALYSIS OF THE LISTENING COMPREHENSION TEST ITEMS: RESULTS 141

7 A PRIORI ANALYSIS OF THE CURRENT ITEMS........ 142
7.1 Items 1-11 142
7.2 Items 25-30 154
7.3 Summary of the content analysis 158

8 ANALYSIS A POSTERIORI: QUANTITATIVE INFORMATION ON THE ITEMS 159
8.1 The test-takers' results 159
8.2 Results of the Rasch analysis of the items 161
8.3 Distractor analysis 163
8.4 Comments on the quantitative analysis 167

9 THE RESEARCH QUESTIONS WITH A FOCUS ON THE INTROSPECTIVE RESPONSES 168
9.1 Description of different types of introspective responses 168
9.2 Consistence between coders for the introspective responses to item 3 172
9.3 A closer investigation of the different introspective response categories 174
9.3.1 Text-based responses: Word-bound responses 174
9.3.2 Text-based responses: Partial comprehension 179
9.3.3 Text-based responses: Résumé 185
9.3.4 Text-based responses: Nonsense 188
9.3.5 Strategy-based responses: Guessing 194
9.3.6 Strategy-based responses: Elimination 208
9.3.7 Metacognitive introspective responses 218
9.4 How do the test-takers' listening and test-taking processes relate to their success in solving the items? 232
9.5 Some problematic features of the individual items 235

10 CONCLUSIONS AND DISCUSSION 258
10.1 What processes are activated and what strategies are employed by the test-takers on seventeen multiple-choice items assessing listening comprehension of French as a foreign language? 258
10.2 How does the nature of the individual MC item influence the employed processes and strategies? 262
10.3 How do the test-takers' listening processes and strategies relate to their success in solving the listening comprehension items? 266

11 DISCUSSIONS ON THE QUALITY OF THE CURRENT RESEARCH STUDY 268

12 IMPLICATIONS OF THE RESULTS OF THE CURRENT STUDY........ 274

12.1 Implications for the development of listening comprehension tests and item writing ... 274
12.2 Pedagogical implications ... 277

13 SUGGESTIONS FOR FURTHER RESEARCH ... 280
TIIVISTELMÄ ... 283
References ... 285
APPENDIX 1 ... 294
APPENDIX 2 ... 301
APPENDIX 3. ... 304
APPENDIX 4: ... 305
APPENDIX 5 ... 308

INTRODUCTION

The comprehension of speech seems to be a fundamental human capacity – what we originally have understood by human communication undeniably involves the two elements of speaking and listening. There are still in the world many people who are not able to read and write, but who use the oral medium for communication. Listening and interpreting parental speech (together with other signals including body language and other paralinguistic features) is the source of learning to communicate for very young infants – they pick up ingredients and experiment with them in their communicative environment. It is mainly only in the context of formal instruction where children learn to communicate through the written medium – within many linguistic traditions with the help of learning to combine the written signs with sounds.

In the learning of second or foreign languages (L2)[1], even if there are contexts where the only medium used may be the written medium – like for example learning to read scientific articles in a foreign language – it is difficult to imagine a context where the oral element is completely missing. However, within the tradition of L2 teaching, the importance set on oral skills and listening comprehension have varied through the decades as a function of the methodology applied. The focus on listening has varied from a proportion near zero to being the main element in the teaching and learning environment (see chapter 1.2).

In a situation where a language is learnt informally in a second language context, it is very often a question of a process of acquisition where the learner listens to the language spoken around him or her, combined with the attempt to read the language around him or her. This learning can be supported in a more formal environment by using more written language and the rules and principles of the targeted language.

Slightly surprisingly, even in 2005, Flowerdew and Miller point out that listening has been treated as the Cinderella of the four language skills. But, as

[1] I will use L2 throughout the text to imply both a second and a foreign language. In this study French is assumed to be a foreign language for all of the present test-takers.

they say, as an essential part of communicative competence, listening is a skill that deserves equal treatment with the others (Flowerdew & Miller 2005: xi).

In the formal learning and teaching context, assessment enters as an ingredient: at different stages of the learning process, the learner, the teacher or other stakeholders want to know how the learning of the individual is proceeding. Different learners have different goals, but in an instructional setting, there is usually a curriculum that forms the norm for what is to be learnt within a particular time span. The assessment instruments vary to a great degree, some may focus on the written language exclusively (reading and writing skills and knowledge of grammatical rules), whereas what would be considered as an assessment instrument covering the individual's entire communicative competence includes at least all the four skills of listening, speaking, reading and writing.

Alderson and Bachman (in Buck 2001: Series Editors' Preface) state that

> [T]he assessment of listening abilities is one of the least understood, least developed and yet one of the most important areas of language testing and assessment. It is all the more curious, therefore, that very little is written in the language assessment literature on the specific constructs, or abilities, that underlie listening, on how to go about designing listening assessment procedures, on how to validate and evaluate listening tests.

There thus seems to be quite a lot left for a researcher to explore.

The framework of my study includes the learning of a foreign language, in a large sense, and within that large framework, the listening skill from a cognitive point of view, and more precisely, the assessment of this skill in a formal, high-stakes context. The methods of exploring these issues combine a main focus of qualitative introspective methods with quantitative information.

Orientation

For the current research work I have three main starting points:

- The nature of the listening comprehension process
- The explicit and implicit problems inherent in the multiple-choice (MC) format
- The validity and validation issue.

The first starting point has to do with the foreign language listening comprehension skill – the construct[2] we are interested in assessing - and its complex and multifaceted nature. By logic and experience as well as through empirical findings we do know something about the comprehension process: a number of

[2] Bachman (2004: 14ff) separates the definition of the construct in a conceptual and an operational part. A conceptual construct definition is generally based on either a theory of language ability, or proficiency, or on the content of an instructional syllabus. The operational definition of the construct gives detailed information about types and numbers of test tasks, the amount of time allowed, and the scoring of the responses to the test tasks.

theories and hypothetical models try to shed some light on the structure of that process. What has been established is that multifarious elements and dimensions are involved. Many details of the nature of this internal and invisible skill are, however, left open to question, and in part we can merely speculate on how the interplay of neurological, cognitive, psychological and social factors is structured. The added dimension of the assessment context further complicates the picture, the question being whether and to what extent the factors involved in a testing context are identical to the factors involved in a real-world target language use situation.

It is essential to point out that the listening comprehension construct should not even be expected to look the same in every assessment context, because there will always be variation as a function of the potential test-takers' targeted level of the skill, as a function of the contents of the text and as a function of the demanded tasks, among others. Some generalities can, however, be established. Buck (1990, 2001) and Rost (2002) point out the necessity of including aspects of language proficiency and comprehension that are unique to listening in the listening construct. Rost (2002: 31, 171-2) mentions various physical, linguistic and psychological features that are typical for spoken language, not reflected in written language and unique to listening - in Buck's (2001: 112) words*:*

> Listening tests ought to require fast, automatic on-line processing of texts that have the linguistic characteristics of typical spoken language.

The construct of listening comprehension skill is furthermore set in the framework of communicative competence/language ability[3] following the principles set forth in language testing contexts primarily by Bachman (1990; Bachman & Palmer 1996) but recognised widely by contemporary language testing scholars. This communicative framework is used by Buck (2001) in establishing a "default" listening framework, a general framework that he recommends should be kept as a basis for the determining of any construct of listening comprehension along the lines of communicative language proficiency, "listening as interpreting meaning in terms of a communicative context" (Buck 2001: 93).

The second building stone of my study concerns the problematic and controversial issue of using the multiple-choice (MC) format for assessing L2 listening comprehension skills. On the one hand, the practicality of the automaticity and the stated objectivity of scoring, together with the possibility of including a relatively large number of test items have made and make the format very popular to use especially in large-scale testing all over the world. On the other hand, the format has been severely criticised, mainly for allowing uninformed

[3] The movement for communicative testing developed primarily in response to the trend towards increased communicative second language teaching. The concept of communicative competence was coined by Hymes in 1966. Canale and Swain's (1980) definition of communicative competence has become canonic in applied linguistics. They define communicative competence in terms of four components: grammatical competence, sociolinguistic competence, discourse competence and strategic competence.

guessing which may distort the reliability of the test scores. From the point of view of validity, the drawbacks of the format concern for instance the possibility of using general problem-solving strategies or test-wiseness to eliminate options on the basis of surface clues. If large-scale tests are also high-stakes tests and the outcome is important for an individual test-taker, this is problematic. How do we (as examiners or users of examination test-scores, as employers for example) know that the score obtained by a test-taker on a test really reflects the level of skill that a test-taker has reached – that is that the results are reliable and valid? The task of creating good MC test items is by no means an easy one, but demands a great deal of knowledge and experience in test construction (Bailey 1998: 131; Alderson 2000: 212; Buck 2001: 142).

In fact, the tester has to assume that it is possible to measure the listening skill in an indirect way, by means of some kind of formal tasks. However, all test formats have their advantages and drawbacks. An important concept here is the trait-method unit, implying that a test-taker's performance is a function of two constant variables: the test-taker's language ability and the test method (See for example Bachman 1990: 225; Yi'an 1998: 21). The underlying assumption is that a test format may influence the processes a test-taker makes use of in a test situation, and the essential question to ask and to investigate is whether these processes are compatible with the particular listening construct that the test developers and users have targeted. In other words: do the results that a test-taker obtains from a particular test of listening comprehension reflect the test-taker's listening comprehension skill in a way intended by the construct?

The third starting point for this study comes out of the two previous angles. When faced with a school leaving examination system where one of the skills to be assessed within the framework of what we can call communicative language ability is named "listening comprehension", all parties involved – from examination system administrators and planners, through test constructors to individual test-takers and other stakeholders – want to make sure that the test given to test-takers and perhaps used as a basis for a general score in a foreign language really measures that skill in a reliable, fair, exact and comprehensive way. We are here concerned with validity and validation issues: making sure that a particular test used in a particular context measures the particular listening comprehension construct that test constructors or examination managers have determined and described. This is indeed the most important question in all language testing (Alderson et al. 1995: 170; AERA 1999: 9; Buck 2001: 1).

As a consequence, the next question concerns the means we have at our disposal for making sure that the construct is in fact measured in a valid way. I have selected as my main approach the qualitative and quantitative analyses of the processes of the test-taker. These invisible and internal cognitive (and affective) test-taker processes can only be reached indirectly, and one method is to let test-takers report on the processes and strategies that they experience during a test event (through some type of verbal protocols). The verbal report method, recommended for example by Bachman (2004: 278) as an indispensable tool for collecting information about test performance as part of the test try-out phase, has been employed in some previous empirical studies exploring the effects of the test format in assessing comprehension skills, and has proved useful for analysing listening tests in the studies by Buck (1991), Ross (1997), Yi'an (1998), Yepes (2001) and Wagner (2006) as well as reading tests in the studies by Nevo (1989) Anderson et al. (1991) and Rupp et al. (2006).

Research questions

In this study I propose to seek answers to three questions derived from the starting points mentioned above; the first being linked to the main aim of the research study, the two following representing related sub questions:

1) What processes are activated and what strategies are employed by the test-takers at seventeen multiple-choice items assessing listening comprehension of French as a foreign language?
2) How does the nature of the individual multiple-choice test items influence the employed processes and strategies?
3) How do the test-takers' listening processes and strategies relate to their success in solving the listening comprehension items?

Data & methodology

The data for my study come from different sources and are of different kinds. For the current study I used parts of the original test of French as a foreign language that was part of the spring 2002 Finnish Matriculation Examination that I modified slightly. The modifications concern the length of the test and the test procedure. Because of the procedure of verbal reporting, I have not included all the original 30 MC items. Altogether seventeen items can be taken to form the MC part of the test. Added to that, I used three texts originally used for MC items as the spoken input for creating four open short-answer questions[4]. The total duration of the test in the research situation needed to be limited to 45 minutes. This was the length of the original test and of a regular Finnish school lesson.

The Matriculation Examination Board undertakes certain statistical analyses, providing the facility level (p-value as proportion of correct responses per item) and the discrimination as well as the attraction of separate options for all test-takers. The proportion of the selection of the key and the two distractors for each item constitutes material for comparison for the seventeen items that I include in the test for the study when *a posteriori* statistical analyses are conducted for the modified test after its use. If the figures are comparable, it can be taken

[4] The open-ended questions are not analyzed in this research study, however, since a focus on and a limit to the MC questions has been found to be more fruitful.

to indicate that even if the test and the test conditions are modified, the test-taking processes and the results that the test-takers obtain can be analysed from the point of view of validity.

An interview conducted with the test developers responsible for the original tests shed light on how a test comes about and explains the conditions and limitations determining the test construction process. Together with information on the current language programme and the school context, the test specifications (in this case mainly in implicit forms) and its regulations, this is important as a practical aspect of test construction and implementation. Practical considerations as well as reflections on the social impact of the test system constitute parts of the important concept of test usefulness, established by Bachman & Palmer (1996)[5].

I administered the current test[6] with 17+4 items within the period of two school years in different (22) upper-secondary schools to 218[7] learners of French on a level corresponding to the target level of the potential test-takers. The task for the test-takers is, besides taking the test of listening comprehension of French as a L2, to provide verbal reports on the test processes. The task is made as simple as possible, by asking the test-takers to justify their option selection at each MC item[8]. These introspective responses form the basis for the analysis of the processes and strategies that emerge at this specific test event.

I first analyse the contents of the spoken texts used as a basis for the test and the MC items (stems and options) related to these texts *a priori* from the point of view of the assumed processing demands set on the test-taker. I speculate on what it is in the text - in terms of phonology, vocabulary, syntax, discourse structure or nature of the contents - that seem to make it relatively easy or difficult to process. I scrutinize the items themselves: what type of information is needed from the spoken text in order to arrive at a correct response: is the information detailed and local or is more global information asked for? Is it necessary to infer and to seek answers from information spread or scattered over longer text passages? The characteristics of the separate items are related to results obtained from earlier empirical studies on the aspects of item difficulty, based on theoretical descriptions of the nature of the listening processes. The characterizations of the separate items provide expectations on how the items may function when given to potential test-takers on the target level of the test. These expectations can be compared with the item analysis obtained after the original administration of the test.

[5] Covering apart from practicality and impact also construct validity, reliability, authenticity and interactiveness

[6] I call the 17 + 4 items used for the current research a test, even if I do not assume that they would represent a reliable measure of the test-takers' entire listening ability. For the students the administered items resemble a test (practice) event.

[7] Of the original 226 test-takers eight test-takers had to be left out from the final analysis as they had already taken some of the test items and this would affect their test-taking processes and distort their test score.

[8] The test form is found in Appendix I and the test-taking procedure is described in detail in chapter 5.2.

The first basic analysis *a posteriori* – after the test administration – consists of a Rasch analysis combined with a graphic distractor analysis; to see how the individual items and the individual options have worked from a quantitative point of view. The test-takers' success in solving the items is taken into account and I use this information for comparing the nature of processing across test-takers with different levels of success.

At this point, the results obtained from the analysis of the contents in combination with the statistical item analysis give indications of the quality of the items, and provide a perspective for the main analysis, which is the study of the introspective responses at each item. The first question in this study is related to the test-takers' processing, and will call for a close investigation of the test-takers' verbal reports (introspective responses): what is the processing like? The approach to this analysis is both qualitative and quantitative. In order to be able to generalize from the total number of 218 x 17 introspective responses, I categorise the responses into nine different text- or task-oriented types. Some broad categories are further specified and divided into subcategories. The basis for this categorisation is data-driven and empirical, mainly drawing on results from pilot studies. I describe the nature of different ways of approaching and solving the listening and test-taking task in the light of the introspective responses. Parallel to that, I approach the differences in these processes both from the point of view of different types of items (the differences being based on their contents or on statistical information) and the test-takers' results on these items.

Organisation of the thesis

In the first part of the thesis I describe the theoretical framework for the study and the state of the art within the particular subfield of cognitive processing approaches to MC item validation. First the cognitive and social nature of L2 listening comprehension is discussed, as well as its role in the L2 teaching context. The complicated and multifaceted skill of listening comprehension through its various processing levels is then treated: from phonology through vocabulary and syntax to discourse and pragmatics. I compare the description with psychological theories on human information processing and Buck's (2001) listening framework. (Chapters 1.1 – 1.5)

I also discuss the notion of listening processes as contrasted with listening or test-taking strategies. Individual listener characteristics are discussed briefly. All this leads to considerations of the established framework as a basis for test construction or validation studies. (Chapters 1.6- 1.10)

Various general key concepts on language testing and assessment are treated and discussed in relation to the construct of listening comprehension. The notion of *usefulness* of a specific test in a specific test situation is a large concept that can be taken to cover all the rest (Bachman & Palmer 1996:18). I discuss the underlying notions and variables of test usefulness in relation to the context of large-scale, high-stakes tests. (Chapters 2.1-2.5)

As test validity and the process of validating a test or items (including the Rasch analysis) are central issues to this study, I contemplate these concepts from a theoretical point of view as well as from the point of view of the context of this test and the current study. (Chapter 2.6-2.7)

The method of collecting information on the validity of test items by means of short written introspection is described within the framework of verbal/think-aloud protocol analysis and the basic principles of human information processing theory. (Chapter 3)

I describe the language examination situation in Finland – including the matriculation examination - in order to shed light on the background context. Interviews conducted with the test constructors serve as a complementary source of background information. The original test procedure is described here as well. (Chapter 4)

In the second part I describe the current study - the main and secondary research questions in terms of the data and the data collection as well as the test procedures (Chapter 5). Next I treat methodological aspects. The principles of the content and statistical analyses that are applied to this test are explained, with a discussion of their limitations in the context of the study. (Chapters 6.1-6.2) I describe the ways of employing the introspective method for the purposes of the current study, as well as the basis for categorising the nine different types of responses that are given by the test-takers as a reaction to the seventeen listening comprehension items (Chapter 6.3).

The third part of the thesis focuses on the results of the (triangulated) analyses. I consider the general contents or characteristics of the test before moving to an investigation of the details of the separate items: what seem to be decisive text and task characteristics influencing the processing demands and the quality of the items (Chapter 7)? The following step concerns the results of the statistical analysis and its implications (Chapter 8).

In chapter 9, which is considered the main chapter and where the most important findings are presented, I reveal the results of the analysis of the introspective responses. The separate types of introspective responses are treated with respect to their nature and frequency at the separate items, as well as with respect to the results of the test-takers. Three variables are considered: item characteristics, the test-takers' success and the test-takers' introspective responses.

In the fourth part of the thesis I draw conclusions from the results obtained in this study, considering them against the background of results from earlier similar studies and the current listening construct (Chapter 10). The qualities and limitations of this type of study are discussed (Chapter 11) with the consequences and applicability of the results of this study to language teaching and language testing (Chapter 12). In the final chapter (13) I focus on future ways to proceed.

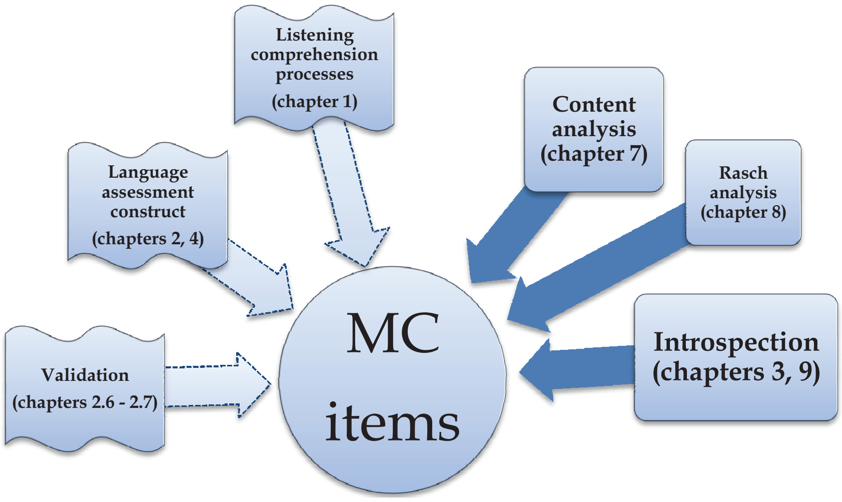

Listening comprehension processes (chapter 1)
Language assessment construct (chapters 2, 4)
Validation (chapters 2.6 - 2.7)
MC items
Content analysis (chapter 7)
Rasch analysis (chapter 8)
Introspection (chapters 3, 9)

FIGURE 1 Overview of the main focus of the thesis

1. THE LISTENING PROCESS

The goal of this chapter is to give an idea of what a general foreign language (L2) listening comprehension process looks like, principally from a cognitive point of view, and first independently of the particularities of the listening event, be it a real-life target language use (TLU) situation[9], a learning situation or a test-taking situation. The process is obviously internal, invisible to the external eye and impossible to get at directly. When taking account of the particular listening context, in addition to the more or less unconscious and automatic processes, there are very often more conscious and deliberate strategies present, both listening strategies and strategies related to the specific listening task (real-world or test-taking tasks).

First, in order to further specify the framework and the limitations of the study, I contrast the cognitive nature of the L2 listening comprehension with the social nature of listening comprehension and language learning in a larger sense. I then present different models of listening comprehension. Some features of the process and of a proposed listening framework (Buck 2001) are focused in more detail, with reference to research results considered important: memory functions, grammatical knowledge (phonology and prosody, vocabulary and schemas, syntax and text organisation,), discourse knowledge (redundancy, elaboration and coherence), pragmatic and sociolinguistic knowledge. This is followed by a description of listening comprehension strategies (as compared to processes, divided into cognitive and metacognitive and with a focus on the strategy of inference). I briefly treat the influence of the test-takers' level of skill and personal characteristics on the processes, with respect to the potential differences between a skilled and a less skilled listener and individual learning styles.

I also discuss the test-taking strategies in the context of a test, as well as the question of the effect of the task. Finally I draw some conclusions as to the implications of the description of the listening processes and strategies on the

[9] For a definition of TLU, see Bachman & Palmer 1996: 44.

current study. I make an attempt to draw a processing model based on the details of the preceding description.

1.1 The cognitive and social nature of L2 listening comprehension

Listening is an activity with many dimensions on many levels. It is a fundamental part of communication in its most basic form, the oral form. One basic characteristic of communication is its interactional character and the social context: there is a purpose for using the language and the purpose is embedded in the individual's social actions at large. This is true both for the first language (or mother tongue) of an individual and for a second or foreign language. The range of languages of an individual are acquired or learnt in a social context, and the purpose for learning in most cases lies in a social context.

The many levels of listening and the comprehension of spoken (foreign) language within an individual have, by necessity, a cognitive as well as a physiological and neurological source. From the distinguishing of sounds, and the interpretation of sound sequences as meaningful units, to the building up of meaning in a larger sense, from lexical units to larger entities, propositions and discourse. The meaningfulness of language, be it spoken or written, however, only exists in the context of use, determined by the participants in social interaction.

The different levels are reflected in the tradition of different models of listening comprehension. The bottom-up model of listening comprehension was developed in the 1940-50s. This model follows a traditional view of communication as transmission of communication, established by Shannon and Weaver (*The mathematical theory of communication,* 1949), where the essential elements consist of a source, an encoder, a channel, a decoder and a destination and the signal is transmitted and received. According to this model, communication can take place without any reference to the speaker, hearer or wider context (Flowerdew & Miller 2005: 25).

The top-down models emphasize the use of previous knowledge in processing a text rather than relying upon the individual sounds and words. The basic assumption is that listeners rely on more than just the acoustic signal to decode a verbal message. Listeners use pre-established patterns of knowledge and discourse structure stored in memory, conceived in a number of ways, as schemas, frames, scripts and scenarios (Flowerdew & Miller 2005: 25). In a parallel or interactive model, the bottom-up and top-down processing is synthesized: language is processed simultaneously at different levels.

These models all rely on a cognitive approach, and many researchers as well as practitioners have found them unsatisfactory when accounting for the complex nature of listening comprehension. Among others, Flowerdew and Miller (2005: 85) propose, added to the "cognitive heart" of listening, further distinct dimensions that they incorporate into a pedagogical model of second language listening. Listening is, they emphasize, also individualized, cross-cultural, social, contextualized, affective, strategic, intertextual and critical (see further descriptions of the model in chapter 1.2).

Within second language acquisition (SLA) research, the issue of a better balance between the cognitive and the social was raised among others by Firth and Wagner (1997). In an article in the Modern Language Journal (1997: 81, 285-300, republished in the Modern Language Journal 91 2007) they examined the prominent view of discourse and communication arguing that the view was individualistic and mechanistic, failing to account for interactional and socio-linguistic dimensions of language. They called for a reconceptualization that would require three major changes in SLA: a) a significantly enhanced awareness of the contextual and interactional dimensions of language use, b) an increased emic (participant-relevant) sensitivity towards fundamental concepts, and c) the broadening of the traditional SLA data base (2007: 758).

In the decade after the Firth and Wagner article, substantial progress has been made in developing models of L2 acquisition that document the impact of social context on the cognitive processes presumed to underlie SLA (Tarone 2007). A table (Table 1) contrasting some of the issues within the cognitivist and social views of SLA is drawn by Larsen-Freeman (2007).

TABLE 1 Cognitivist and social views of SLA according to Larsen-Freeman (2007)

	Cognitivist SLA	Social SLA
1. Role of Context	Social context is the site in which L2 acquisition takes place; however, if you change the context, the acquisition place remains the same. The goal is to search for universals that transcend individual contexts.	Social context influences performance. Social factors are related to systematic variation in learner language. Each context is unique although certain generalizations, such as turn-taking principles or observations about repair, can be made.
2. Nature of Language	Language is a mental construct.	Language is a social construct.
3. Nature of learning	Change in mental state	Change in social participation
4. Primary Research Focus	The primary focus is on language acquisition (how people learn a language, not how they use it). Given this focus, what is important are cognitive factors of knowledge representation, processing and recall.	The primary focus is on language use. Language use and acquisition cannot be easily separated. Therefore, what are important are social/interactional factors and their effect on the language used.
5. Objects of Inquiry in Language-Focused Research	What is of interest is the aggregation and increasing complexity and control of linguistic structures by learners.	What are of interest are discursive routines of communication processes. There is also a need to look at the purpose of talk; a fundamental perspective to language is most helpful.

6. Identity of Research Participants	The salient identity of the participant in a research study is that of a learner.	The identity that the research participant adopts makes a huge difference, and it may not be that of a learner. For example, in the moment, a learner may not "perform his or her competence" because he or she might want to align socially with another less competent peer.
7. Perspective on Evaluating Learners' Progress	Progress is measured by where along the route toward target proficiency the learner is as indicated by the learner's linguistic performance.	What is at issue is what the learner does with the resources that are available. Look at what the learner does to get his or her message across, not what the learner cannot do.
8. End state	The end state occurs when learner language and target language are congruent or where learner language is stabilized/fossilized.	There is no end state.
9. Philosophical Orientation	Scientific, value-free inquiry	A critical view
	Modernist	Postmodernist
10. Research Site	Varied, sometimes natural environments, sometimes experimental, where data are elicited	Varied contexts where language is used naturally and heterogeneously
11. Primary Level of Research Conceptualizations	*Macrolevel* idealizations, in other words, native speaker, learner	*Microlevel* social relationships that are being achieved through talk in progress
12. Attitudes Toward Acceptance of SLA Theories	One theory will prevail; empiricism will determine which.	Multiple theories are welcome, even necessary. Relativist; pluralist
	Positivist	

Larsen-Freeman (2007: 783-84) speaks for the fruitfulness of the intersection of the two perspectives which seems to represent the chaos/complexity theory point of view, where

> [...] what endures is not a rule-based competence, but a structured network of dynamic language-using patterns with specific information about instances of use retained in the representation. Because these variegated language-using patterns emerge from language use, they are not only characterized by linguistic features, but they are also sometimes accompanied by gesture, unique prosodics, and by affective, cognitive, and episodic associations, experienced as they are embedded in a sociohistorical context.

It seems logic that language use and learning is thus shaped by and dependent on both the cognitive capacity of the individual and the social environment in which this capacity is put into use.

How does the assessment of second language or foreign language (L2) listening comprehension relate to this view? There are several issues at stake. First of all, we have to ask whether the classroom-based language learning event should be considered a socially constructed target language use event. It certainly is one language use situation among other innumerable situations, with some traits that, however, makes it largely different from "authentic" or genuine language use situations with native speakers. The purpose of using the target language in class may remain unnatural, in case the participants (the learners) share a common first language. The purpose for learning a language is, however, socially determined[10]. The goal of learning a target language is, in most cases, to be able to communicate, in the large sense of the word, in the target language: to be able to use it in order to gain citizenship, be able to work, study, travel or practice a particular hobby or simply interact with people and ideas from a different culture. The interactional point of view is always, at least implicitly, present.

Second we may ask whether a language testing event is a language use event. Again, it certainly is an occasion to use the target language. Added to that, the reason to participate in a language test is socially determined. A test-taker may have to prove his or her ability in order to gain a grade (in the school-leaving context) or to be compared with peers (in the entrance examination context), or in order to fulfil some criteria in a summative testing context. In a diagnostic assessment situation, the outcome can affect the socially related concrete teaching or learning context, as for example the placement in a learning group. Even the desire to prove to oneself that a certain language proficiency level has been reached can be said to be socially determined. However, the "micro-level" purpose of the set task can in many cases be unnatural and artificial, lacking the important communicative, interactional aspect present in many natural "authentic" target language use situations.

Another question is related to the view we have of language, which shapes the way the construct to be assessed is described. In her comparison of the cognitivist and social views on language, Larsen-Freeman (2007) points out that in the cognitivist view, the learner's performance is measured with a target level proficiency in mind. The focus lies largely on the lacking bits of proficiency and on failures. In the social view, what counts is the learner's capacity to adjust to the situation – which seems to be related to a "metacapacity" of knowing how to use resources – an interlocutor, for example - in a particular language use context.

When we want to assess L2 listening comprehension, the first question is related to whether listening comprehension is a skill that should be measured in isolation. This, again, depends on the construct, but in a communicative framework, what seems to be relevant is the functional use of the skill, combined with whatever other skills may be at stake in a particular language use context. However, in most summative, norm-referenced language assessment contexts, the skill of listening comprehension is assessed separately – without being fully isolated, since most tasks, by necessity, include some other skill for proving comprehension – and individually, without the possibility of using peers, inter-

[10] In the large sense of the word, including also for example the educational, political and socio-economical aspects.

locutors or other interactional resources in the situation. We appear to be far from the socially shaped language use situation, when we focus on the cognitive capacities that we hope to assess in a valid and reliable way.

It is difficult to imagine a socio-culturally determined large-scale and high-stakes assessment context, where the individual's social and interactional capacity would be one central dimension of evaluation. In the classroom context, where the focus may also be on diagnostic assessment, the situation is different. Dynamic assessment is described by Poehner & Lantoft (2005) as a pedagogic approach grounded in a specific theory of mind and mental development with the defining characteristic of negotiation of mediation aimed at development. This mediation must be tuned to those abilities that are maturing, and as they mature further, the mediation must be continually renegotiated. The key concept is the zone of proximal development (ZPD), according to a theory originally proposed by Vygotsky. Dynamic assessment aims at supporting long-term development. Assessment and instruction are considered inseparable, as they form a necessary unit for learner development. Interaction during the administration of an assessment is an indispensable component of the procedure (Poehner & Lantoft 2005: 261).

> "The principle underlying dynamic assessment is that a full picture of what an individual or group is capable of does not emerge unless and until the ability is not only observed in independent performance but is also pushed forward through specific forms of intervention and/or social interaction between learners and assessors. Thus, DA represents a perspective on assessment and instruction in which these are seen as two sides of the same coin".

It is difficult to imagine this sort of intervention within assessment contexts following traditional views of assessment where reliability and validity of the interpretation of test scores are essential.

Takala (1998), in an article discussing recent developments and persistent dilemmas within the field of language testing, gives an interesting overview comparing the relative advantages and disadvantages between authentic, or alternative, or performance assessment[11] and traditional assessment (see further chapter 1.9 below). Takala (1998) speaks of a trade-off where advantages are bought at the expense of disadvantages. The importance lies in making informed choices, depending on the particular assessment context at hand. In analogy with Larsen-Freeman's idea of the fruitfulness of the intersection of the socio-constructivist and cognitive views, the idea of combining the traditional assessment methods with alternatives methods, in an attempt to maximise the advantages of the two, may be a solution. For the Finnish assessment context, for example, Takala[12] envisions the advantage of a combination/coordination of the evaluation conducted by the language learners' teacher and an external evaluation (like the Matriculation Examination). The teachers should do more diagnostic assessment of the learners, in order to gain an insight into their strong and weak processes and conduct more observation and interpretation

[11] See also for example Bachman 2000: 11-15 and Messick (1994).

[12] Personal correspondence 17.6 2008.

with fewer tests in the traditional sense. Takala suggests having teachers carry out speaking and writing assessment tasks, following *CEFR*[13]-adapted scales and criteria. This bears some similarity to the idea of having portfolios. The external evaluation/assessment would stand for 60 to 75 % of the final grade of a learner. Takala points out, however, that much more teacher training in assessment literacy, numeracy and practice is needed in order to carry out such reforms.

Within a recent research project, ToLP[14], on Finnish L1 (as a first language) and L2 literacy practices, the assessment practices and especially the student's own role in the practices were studied. According to the results of a survey, the assessment carried out is experienced, by both the teacher and the students (9th graders), to be very teacher-centred, as is the grading of the students. Assessment is conducted - especially it is experienced by the students to be conducted – mainly at the end of a learning period. Students seem to trust the teachers' grading - 73% of the students' agreed with the statement *'My foreign language grade gives an accurate picture of my achievement'*. Neither teachers nor students identified any other important actors in assessment. Interestingly, teachers reported that self- and peer-assessment was conducted more often than the students did. The reasons, the researchers speculate, may be that the students do not always know what activities include self- or peer-assessment. On the other hand, the teachers may be conscious of self- and peer-assessment as parts of modern language assessment, and tend to exaggerate their use of it in the survey. The researchers conclude that the assessment practices in the Finnish comprehensive school are partly consistent with the national curriculum. Self-assessment and peer-assessment have gained ground and are used to some extent, although teachers and students tell a somewhat different story about this. The researchers are worried about the fact that over a third of the students reported that self-assessment was done never or only rarely (Huhta & Tarnanen 2009).

In the context of the current study, where the focus is on one part of the Finnish Matriculation Examination (as it looked like in 2002), a high-stakes and large-scale assessment instrument, the cognitive view prevails, reflecting the principles underlying the current assessment context of the object of study, classified within the traditional assessment paradigm. The methodology used in the study, based on some principles of psychometrics and verbal protocol analysis, also principally follows the cognitive view. Nevertheless, I am aware of the much larger socio-cultural and pedagogical context that lies behind all language use and learning.

[13] *Common European Framework of Reference for Languages: , Learning, Teaching, Assassment* . CUP, Cambridge. *http://www.coe.int/T/DG4/Portfolio/documents/Framework_EN.pdf . See overall description in Appendix V.*

[14] ToLP, *Towards Future Literacy Pedagogies. Finnish 9th graders' and teachers' literacy practices in school and out-of-school contexts,* is a multidisciplinary research project (2006–2009), carried out at the University of Jyväskylä, Centre for Applied Language Studies. www.jyu.fi/tolp

1.2 The role of listening comprehension in L2 teaching

In order to place the construct of the skill of listening comprehension into perspective, I will give a brief overview of the role of listening comprehension in L2 teaching. Flowerdew and Miller (2005: 3-20) report on the development of the role of listening comprehension through the different language teaching approaches: the grammar-translation approach, the direct-method approach, the grammar approach, the audio-lingual approach, the discrete-item approach, the communicative approach, the task-based approach, the learner-strategy approach and the integrated approach. Language teaching methods initially did not recognize the need to teach listening, but subsequent approaches used a variety of techniques to develop specific or general listening skills (Flowerdew & Miller 2005: 19).

In the grammar-translation method, where the main goal was to learn to read (Latin and Greek) literature, listening did not have a role at all. The reaction to this lead to the direct method, where the idea was that learners can best learn by means of an aural /oral method with the exclusive use of L2. The listening skills were first in focus and other skills came later.

The grammar approach had at its main focus analysis of the language by its components. During a listening exercise according to the principles of this approach, students usually look at a written text while listening – the task being to figure out the meaning and structure of the text. It is a very classroom-oriented approach with little relevance to real-world language use (Flowerdew & Miller 2005: 7).

The audio-lingual approach was generated by the U.S. Defense Forces language program at the time of World War II. The core of listening consists of a focus on imitating pronunciation and grammatical forms by means of drills and exercises. The students listen to a tape recording or to their teacher, and then record their own version in creating "good habits", minimizing interference from the L1. This method became very popular with the arrival of the language laboratories.

In the discrete-item approach the teaching deals with individual vowel and consonant sounds, with stress and tone. They are presented and drilled, compared and contrasted. The exercises practicing listening are isolation-type tasks with processing on a discrete-item basis.

The communicative approach radically contrasts to the approaches above. Its value comes from the link to real-life activities, to the functional purposes of the listening activity. This approach is said to "*look at what people do with language and how they respond to what they hear*" (Flowerdew & Miller 2005: 12). The activities should be useful to the students, operate above the sentence level and involve actions. Mistakes are tolerated as long as they do not interfere with the communication.

A typical activity within the task-based approach would be to listen to "authentic" situations and transfer the spoken information to a graphic form.

Language spoken at normal speed has the typical features of accents, hesitations, fillers and ellipses and the outcome of the task is unpredictable. It is the result of an "*interaction between the task and the task situation*" (Flowerdew & Miller 2005: 14) and the students' holistic inferential strategies.

In the learner-strategy approach the learner is given a more active role in order to develop an awareness of skills related to listening. The learner should have the opportunity to experience listening strategies in different contexts and for a variety of reasons. The idea is also to find meaningful pre-listening activities and to share the awareness of listening strategies with fellow learners.

As is pointed out by Flowerdew and Miller (2005: 18), teachers of today use different approaches to teaching listening according to their knowledge and preferences. Textbooks have their influence, and depending on the target level of the learners, they offer exercises ranging from those focusing on more traditional features (listening for details or for gist) to exercises with the goal of developing a range of listening strategies.

1.3 Models of the listening comprehension process

Even if listening comprehension can be seen as a cognitive procedure where meaning is to be extracted out of spoken input, it contains much more than just the decoding of the auditory signal. Rather, the listening process is a process of inference covering both linguistic and non-linguistic knowledge within the listener: listening comprehension consists of phonological, lexical and syntactic knowledge about the language, as well as a competence needed to interpret the text. Due to the demands of the listening context the process has to occur automatically and in real-time and is linked to the listener's background knowledge and dependent on his or her individual characteristics.

A framework for studying thinking is outlined by cognitive psychologists Ericsson and Simon (1987). In their framework the base consists of an information processing theory of human cognition. The assumption that human cognition is information processing implies that the processes can be seen as "a sequence of internal states successively transformed by a series of information processes" (Ericsson & Simon 1987: 25). Information is stored in several memories including sensory stores, short-term memory and long-term memory.

Importantly, researchers have long since rejected the first assumptions of the comprehension model where the building up of the meaning of an utterance is linear, proceeding from the smallest unit, the phoneme, to the morpheme, the lexeme and the phrase, finally arriving at an interpretation of the utterance as a whole. Psycholinguists call these first linear models of the information processing models bottom-up processes (see Rost 1994: 4). These are obviously far too limited to account for the entire linguistic process. A great deal of information is in fact built up within the listener deriving from his or her interests, values, attitudes, motivation and background knowledge. The use of these essential complementary sources is called a top-down process. Reading theorists speak about interactive processes, where various types of knowledge are used in any order or simultaneously, interacting and influencing each other (Buck 2001: 3). Similarly, Greene (1986: 96) describes a *heterarchical* "cooperative" model, where there are interactions between different types of knowledge in language processing.

The relative use of bottom-up and top-down processes in listening comprehension has been investigated by Tsui & Fullilove (1998). They point out that the results obtained through earlier research are not uniform as to which processes are decisive for successful listening comprehension. Some researchers claim that skilful listeners know how to make use of their background knowledge and schemas for forming hypotheses to help the interpretation of the spoken text (a focus on the top-down process). Others believe that the main skill is the ability to treat the linguistic details as fast and as efficiently as possible (a focus on the bottom-up process). According to research conducted by Tsui and Fullilove and based on language examination in Hong-Kong, the essential ability that discriminates between weak and strong test-takers appears to be the combined capacity of treating the input efficiently and of interpreting the message correctly. In accordance with previous studies, their findings indicate that skilful learners shift between the two processes – using the acoustic and linguistic textual information to verify or to reject the formed hypotheses about the meaning. Text processing is thus a truly interactive process.

It has been assumed (in analogy with studies on reading comprehension, see Carrell 1988) that less skilled listeners rely excessively on one or the other of the processes. Excessive use of the top-down process leads to a situation where the listener is unable to verify the hypotheses by the linguistic details in the text. On the other hand, overuse of the bottom-up processes implies that the learner doesn't make use of his or her background-information for the interpretation of the text.

There are several theoretical models of the listening comprehension process that include similar stages. Three models are summarized in Table 2, together with difficulties associated with each of the stages.

TABLE 2 Models of the listening comprehension process and difficulties associated with each stage

Models of the listening comprehension process			**Difficulties**	
Anderson's (1985) three-phase comprehension model	**Brown's (1995) four-stage model**	**Turner's (1995) mental processes involved in listening comprehension**	**Dickinson (1987)**	**Goh (2000)**
Perceptual processing	Identifying stage	Taking in a "stream of sound" Organisation into segments or chunks in the echoic memory	Discrimination Segmentation Speed of delivery	No recognition of known words Neglecting subsequent parts of text Not chunking streams of speech Missing beginning Concentration problems
Parsing	Memory searched	Short-term memory: inspection, relationships, retention and rejection	Vocabulary Short-term memory	Forgetting of heard text No mental representation formed No understanding of subsequent parts
Utilization	Storing information	Review with information & possible storage in long-term memory	Compensation Grammar Figurative language Communicative value Reference to extratextual & implicit facts Cohesion & coherence Discourse organisation Long-term memory	No understanding of message No identification of key message
	(Information put into use and enacted upon)			

Anderson's three-phase comprehension model (1985) divides human information processing into three stages: perceptual processing (attention focused on the oral text and the sounds are retained in echoic memory), parsing (forming propositional representations that are abstractions of the original message) and utilization (relating a mental representation of the text meaning to existing knowledge – thereby enhancing comprehension and, most likely, retention of the presented information). This model has been frequently used in studies of listening strategies (e.g. O'Malley et al. 1989; Goh 2000; 2002; Vandergrift 2003).

In the four-stage model of a process of understanding described by Brown (1995: 61), the fourth stage is said to be optional. Here the first stage is an identifying stage, where the information expressed in the text is identified. During the second stage, existing files in memory are searched in trying to relate the new information to old information. The third stage is about storing information in memory, cross-referencing to relevant existing files or setting up a new file for new information. During the fourth stage, the information is put to use and acted upon.

Turner (1995: 5) proposes the following mental processes involved in understanding the spoken word. The listener

- Takes in a "stream of sound"
- Attempts to organise it into segments or chunks in the echoic memory
- Holds on to the units of meaning in the short-term memory and makes more detailed inspection, seeking relationships between units, rejecting what seems redundant and holding on to what seems relevant;
- Reviews what (s)he hears in the light of what is known by reference to the information held in long-term memory;
- Continues to take in more information through the ears;
- Stores the meaning of what (s)he has heard (not the actual words) in the long-term memory if it seems appropriate.

These processes take place simultaneously, not sequentially.

In Chamot's (1995) summary, language comprehension is viewed by cognitive theory as an active process in which meaning is constructed through a complex interaction between the characteristics of the input, the types of declarative knowledge that are accessed, and the use of strategic processes to enhance understanding (Chamot 1995:16).

Flowerdew and Miller (2005), in their pedagogical model for second language listening, place at its heart the cognitive models of listening (bottom-up, top-down or interactive processing) but complete this with other dimensions of listening, giving it a more complex structure. The added dimensions are individualized, cross-cultural, social, contextualized, affective, strategic, intertextual and critical (Flowerdew and Miller 2005: 85-95). Individual variation refers to differences in proficiency stages, text and task types as well as learning styles. Flowerdew and Miller (2005) point out that some individuals may prefer focusing on bottom-up processing, whereas others emphasize top-down processing. Generally, beginners need more time in developing basic bottom-up skills (as for instance decoding), whereas more advanced learners need to develop top-down skills (e.g. applying schematic knowledge).

The cross-cultural dimension refers to the tendency for different cultures to give rise to different schemata and different expectations and interpretations of a given text (Flowerdew & Miller 2005: 87). If, the researchers point out, culture is taken in its broad definition, it can also embrace the differences in age, gender, social and professional position, attitudes, values, beliefs, and general world knowledge. Thus we have a spectrum of individually and culturally determined factors that influence the individual interpretations of a spoken text.

The social dimension has to do with the typical real-world communicative listening situation: the dialogue or the conversation. Listening is by nature a social activity – not just a psychoperceptual process – where both the speaker and the listener affect the nature of the message and its interpretation. A conversational situation also may include side participants, the overhearers, who have their own roles. Flowerdew and Miller (2005: 89) mention the fact that very often second language learners are in fact overhearers, listening to recordings of conversations or monologues, having no active role and being unable to participate or to guide the interaction.

The multidimensional aspect of listening also includes the fact that the meanings derived from listening are affected by relations between the topical particular text with other texts or contextual, parallel activities to the listening situation proper (Flowerdew & Miller 2005:90). This contextualized dimension is also natural in a real-life communicative situation but perhaps less so in the language learning and testing context, as is pointed out by the researchers. This dimension seems closely related to what Flowerdew and Miller (2005) call the intertextual dimension, where they discern a broader type of textual relation, linked to conceptual knowledge as well as genre and register. Intertextuality is paralleled with the concept of schemata, and it is pointed out that this aspect of comprehension demands a high level of familiarity with the target culture (ibid, 94).

Affective variables influence the comprehension process in that attitudes, motivation, affect and physical feeling are prerequisites for a decision to listen. This is, according to Flowerdew and Miller (2005: 92), particularly important for the learning context. The strategic dimension has to do with metacognitive ability: the learners' monitoring of their use of language. From a learning perspective, the learners' identification of their preferred strategies is likely to enhance their acquisition of the listening ability.

The critical dimension involves interpreting language critically in the light of unequal distribution of power, as it is the powerful members of society who tend to control the setting, participants, topics, style, rhetorical patterning and interaction in which discourse is created (ibid: 95). The task of analysing a text critically seems to be a high-level task for L2 learners, but the authors point out that from a linguistic point of view, the utterances are not necessarily that complex, and could well be used at more elementary levels of comprehension.

Judging by the theoretical frameworks and research results presented above, it is obvious that the listening process is anything but simple. Dickinson's (1987) list of potential difficulties encountered at various stages of the listening process provides us with an idea of just how complicated the process is and what problems a learner may experience. These problems can be related to (see Table 2):

- Discrimination: not being able to perceive differences between sounds
- Segmentation: being incapable of identifying sound sequences as words.
- Vocabulary: not knowing the meaning of the words that are heard.

- Short-term memory: not remembering the heard words long enough to be able to form hypotheses on their meaning.
- Compensation: being unable to guess the meaning of unfamiliar or unidentified words.
- Grammar: not understanding the meaning of expressions with certain structures or word orders.
- Figurative language: being unable to understand the non-literal meaning of the language (metaphors, idiomatic expressions).
- The communicative value of the discourse: being unaware of the added and hidden or different meaning of an expression
- References to something outside the text: not understanding references to facts that are not explicitly mentioned by the speaker.
- Cohesion and coherence: being incapable of following the semantic thread from one phrase to another.
- Discourse organization: being unable to understand the structural discourse markings.
- Long-term memory: not being able to understand enough of the heard text in order to understand the whole.
- Speed of delivery: being unable to follow the text because of too fast a speech rate.

Even for native speakers, but especially for language learners, speech accents can be problematic. The accents can differ from the standard both due to geographic and social distances, which is true to most languages. In fact, an unfamiliar accent can make comprehension almost impossible for the L2 listener (Buck 2001: 34-35).

These problems can be compared with the very similar ones that have been empirically found and that are related to the three processing stages of perception, parsing and utilization. In a study conducted by Goh (2000), during the perception phase, listeners did not recognize words they knew; they neglected the next part of a text when thinking about meaning; they did not chunk streams of speech; they missed the beginning of texts and had concentration problems. During parsing, listeners tended to quickly forget what was heard, they were not able to form a mental representation from words they heard; they did not understand subsequent parts because of earlier problems. At the utilization phase, some listeners did understand the words but not the message or did not identify key ideas in the message (Goh 2000:59) (cf. Table 2).

1.4 Cognitive factors related to the L2 listening process

I will proceed by going through some of the elements and dimensions treated above in more detail, reviewing such studies on some of these details that seem particularly relevant in the context of my study: I will focus on cognitive, intra-individual factors related to the foreign language listening process. First, I will outline a basic idea of how the memory system works in treating the incoming information. Next, I will go through what our long-term memory stores include in terms of (declarative) knowledge about the language – in general terms and specifically concerning the target, foreign language (including passages on grammatical knowledge – phonology and prosody, vocabulary and schemas, syntax and text organization -, on discourse knowledge – redundancy, elaboration and coherence, – and on pragmatic and sociolinguistic knowledge). Finally, I will treat the more conscious language use and comprehension strategies that are necessary ingredients in processing language along with some general inferencing and reasoning abilities.

Memory functions

Memory is crucial for the listening comprehension ability and thus a necessary ingredient in the interactive listening process. It consists of a complex procedure based on transformation and use of perceptions and sensory experiences. A listener can manage to handle all sensory impressions he or she obtains by creating internal representations of the surrounding world so that he or she can deal with the information efficiently (Noblitt 1995). When the sensory impressions reach the auditory cortex of the brain, the listener's attention is directed to the formulation of a representation or a mental model of what is heard in form of symbols. Then specific parts of this representation are focused in order to permit updating of what we already know according to the new incoming information (Rost 1994: 66-67). One important aspect of language understanding is the integration of the new information in an utterance with the overall information in the speech act. In other words, it is essential to be able to separate information already known from new facts. We experience information overload when there is too much new information in discourse.

Short-term (or working) memory only seems to be able to handle a limited amount of ideas or propositions containing new information: two to four propositions seems to be the optimal amount. If there are too many propositions with new information it is difficult for the listener to obtain a feeling of coherence in the text. In a situation containing too much information, the listener has to select the parts that he or she will pay attention to (Rost 1994: 70). However, the more meaningful the relationship between the propositions, the more a listener can hold on to (Greene 1986: 6).

It is essential to forget the irrelevant or redundant information in order for the relevant ideas to be stored and integrated in memory. Very often it is not a case of complete repression, but of placing redundant information in the background. When needed and when an effort is made, it is possible to recall this information in memory, depending on the time that has passed since the storing event.

Long-term memory contains a network of images and representations that are formed when new experiences in the form of external stimuli are gained. The sensory images are transformed into memory codes, which are simplified forms that give a person the possibility to reach and to reconstruct her previous experiences.

It takes time to find and to activate the necessary memory links and to sort out the elements in a listening situation after the presentation of a new stimulus (Rost 1994:71). It should be considered that the time available in short-term memory to decide what to do with an incoming message is less than 15 seconds (Flowerdew & Miller 2005: 24). Especially in a foreign language, the pressure set on the memory processes is heavy, because the available memory resources (time and cognitive capacity) to sort out the input are very limited compared to the heavy load from the incoming input.

Long-term memory can only process a limited number of new pieces of information. The human ability to take in new facts, ideas and feelings in long-term memory depends on their relevance for the reorganization of memories, and the effort put into memorizing facts. The process of memorization can be categorized according to its type: textual memory (to remember exactly what has been said), organizational memory (to recall in memory an overall structure of an event), memory through inference or interpretation (to remember consequences or ideas) and evaluative memory (to remember one's affective reaction in a situation).

Central factors that may complicate the process of listening and recall include: lack of attention and interest, information overload, inability to organize information, tendency to distort facts and lack of relevance in the act of recalling (Rost 1994:69). Therefore, in order to be able to recall what we listen to, we must have: proper attention and interest, an appropriate amount of new information, opportunity to organize and rehearse as well as a relevant incentive.

The difference between automatic and controlled processes is interesting especially in the context of second language comprehension. Automatic processes require sufficient training to develop and this training is provided by controlled processing (Nagle & Sanders 1986: 16). Controlled processing occurs in performing new language tasks, which require a high degree of focal attention, and is therefore associated with many foreign/second language comprehension situations. The development of automatic processing is critical to comprehension because too much controlled processing may lead to overload and breakdown. The lower the proficiency level of the listener, the more the processes have to be controlled, simply because of a lack of sufficient exposure or training for the processing to have become automatic. This is related to the speech rate, the amount of spoken input that the listener has to treat at one listening event, as well as to the context and the language use task at hand.

1.5 A listening framework for the listening construct

For the purpose of establishing a formal construct - a description of the ability or skill that is to be measured and that is to be operationalized in the shape of an assessment instrument - Buck (2001) has - on the basis of the general description of language ability by Bachman & Palmer (1996) - created a framework covering the main factors or subskills of the listening comprehension skill. This framework includes the main ingredients or factors related to the listening process (see Table 3).

TABLE 3 A framework for describing listening ability (adapted from Buck 2001:104)

Language competence	- Both fully automated procedural knowledge and controlled declarative knowledge
• Grammatical knowledge	- Understanding short utterances on a literal, semantic level - Phonology, stress, intonation, spoken vocabulary, spoken syntax
• Discourse knowledge	- Understanding longer utterances or interactive discourse between two or more speakers. - Knowledge of discourse features: cohesion, rhetorical schemata, and the structure of unplanned discourse…
• Pragmatic knowledge	- Understanding the function or the illocutionary force of an utterance or a longer text: understanding what an utterance is intended to do, and understanding indirect speech acts and pragmatic implications
• Sociolinguistic knowledge	- Understanding the language of particular socio-cultural settings and particular contexts: appropriate linguistic forms and conventions – slang, idiomatic expressions, cultural references, and levels of formality…
Strategic competence	Executive processes that fulfil the management function in listening; the ability to use language competence
• Cognitive strategies	Mental activities related to comprehending and storing input in memory
Comprehension processes	Processing of linguistic and non-linguistic input
Storing and memory processes	Storing input in working memory or long-term memory
Using and retrieval processes	Accessing memory
• Metacognitive strategies	Activities performing an executive function in the management of cognitive strategies
Assessing the situation	Taking stock of conditions surrounding a language task: assessing one's own knowledge, the available resources and the constraints of the situation before the task
Monitoring	Determining the effectiveness of one's performance while engaged in a task
Self-evaluating	Determining the effectiveness of one's performance afterwards
Self-testing	Testing one-self to determine the effectiveness of one's own language use or the lack thereof.

In the following are further discussed specific aspects of this framework, pertinent for the context of the current study, where the L2 context and the target ability level shape and limit the construct.

1.5.1 Grammatical knowledge in listening – processing and problems

Following the framework above (based on Buck 2001) the first component of language knowledge is labelled grammatical knowledge. It consists of all elements included in the understanding of short utterances on a literal semantic level. I will discuss these elements under the three headings of phonology and prosody; vocabulary and schemas; and syntax and text organisation.

Phonology and prosody

According to Nagle and Sanders' (1986) schematic description of the listening comprehension process, the sensory register catches the input in the shape of sound images that are transported to short-term memory, where the sound is divided into units carrying meaning (words, utterances) according to the information and the knowledge that is stored in long-term memory. The landscape model *(modèle paysagiste)* for the reception of language signals, developed by Lhote (1995), that emphasizes the importance of the sound image, takes into account variations of the different units in the landscape in linguistic information exchanges (voices, sounds, rhythm, intonation, tone, silence). Every language has its own sound landscapes that the listeners reconstruct in their own way, parting from the mental representation as a whole. The difficulty in recognizing the acoustic forms in a foreign language, says Lhote (1995), is the greatest obstacle to overcome in the perception of this language. When listening to French, learners should focus their attention on the end of a rhythmic group. A rhythmic group is a natural mechanism for the organization of discourse, structuring the syntax into *thema* and *rhema*. When a sentence or a group of sentences is pronounced, the words are grouped to form a meaningful entity. The size of this entity varies, but is rarely more than nine syllables, because of restrictions in memorizing capacity. A strong connection ties the lexical units within the rhythmic group together. *Liaisons* (contractions) and *elisions* (reductions) contribute to the cohesion of the elements within the group. This might lead to the impression for the learner that the group consists of one single word. An important part of the difficulties related to the treatment of utterances is related to the fact that the flow of speech for the L2 learner seems continuous, without clear borders between meaningful units. There does not seem to be time enough to analyze the current input, before the next input presents itself (Noblitt 1995).

The phonemic indetermination of the sound signal and the reasons for it are discussed by Dirven and Oakshott-Taylor (1984). First, the phonemic form of separate words is often obscured by the phonemic rules for reduction, assimilation, and elision (called *sandhi*- variation). The difficulties experienced by the listener are caused by the language specificity of these rules. In French there are rules for the *liaison* according to which an adjective or a pronominal determinant is linked to the following main word: *les‿hommes, mon‿ami, deux‿ans*. There are also bindings between a preposition and the following word: chez‿*elle,* between a personal pronoun as a subject and the verb: *nous‿arrivons,* between an adverb and the adjective *très‿heureux,* as well as in fixed expressions *tout‿à‿fait* (Malmberg 1976: 142-44). Another problematic phenomenon in French is the so-called *∂ instable* that, taken together with the speech rate, often causes problems for, for example, Swedish-speakers (Tegelberg 1995: 78-79). *∂ instable* has no correspondence in Swedish or in Finnish, and its name comes from the fact that it is sometimes pronounced, sometimes not. The speech rate and the style level are decisive in determining whether or not it is pronounced: in careful language or at emphasis it is more often pronounced. Another factor determining the pronunciation of *∂ instable* is its placement in the rhythmic group. Thus is the pronunciation of a word highly dependent on its surroundings.

The co-articulation of adjacently occurring sounds makes the distinction of the phonemes difficult. One of the consequences of blurred phonemic discrimination is a high frequency of ambiguities in continuous speech. The amount of homophones in continuous speech is much greater than phonology on the word level would indicate. As Henrichsen (1984) has found out in her study, non-native listeners have relatively more difficulties understanding texts with contractions and reductions than natives have.

Everyday expressions are examples of sound sequences that are difficult to understand in context, because of phonemic assimilation and reduction (Ur 1984). In other words, the difficulty in perceiving these colloquial expressions is caused by the natives' fast and not always very clear pronunciation.

The learner is, however, supported in his interpretation of the meaning of the text by the ability to correlate the speech act with its prosodic features. Consequently, the stress patterns of words and phrases, along with the intonation, are important for understanding. Lynch (1998) points out that the prosodic features have a direct influence on the listener's way of dividing up and interpreting the units in an utterance. Prosody helps dividing a spoken text into units of different sizes, and the prosodic continuity supports the sound signal through coherent modulations of tone and volume in relation to possible structures. The redundant indications of the melodic structure that accompany the utterance help the treatment of the spoken language to a great extent. Intonation often correlates with silence, pauses and variations in the speech rate, and these factors add to the probability that the speaker will reach his or her communicative aim, i.e. make the listener understand.

Pauses and fillers such as *euh, ben, bon* are interpreted as features that aid comprehension, probably because they allow more time for treating an utterance. Nevertheless, some studies that Rubin (1994) mentions show contradictory results that make the researchers consider the effect of the competence level of the listener. The pauses (and especially the filled pauses) may support the comprehension process for listeners at an advanced level, while the treatment of the text becomes more complicated for the weaker listeners. Contrary to hypotheses made by Freedle and Kostin (1999) of the relative difficulty of items associated with texts with long or filled pauses compared to texts with fewer pauses, the latter ones turned out to be less difficult. The researchers' conclusion was that any type of interruption in the reception of the speaker's ideas causes complications for the text treatment process.

Research results seem to indicate that there is a high correlation between speech rate and comprehension problems. The experience of a spoken text having an excessive speech rate is related to the lack of automated interpretation processes. Buck (2001: 40-41) assumes that other variables, related to the vocabulary, the syntax and the theme, together with the speech rate, affect the difficulty of the interpretation of a spoken text.

Deficient segmental information derived from the acoustic signal is usually compensated by the listener's overall knowledge of the language and by the expectations raised in the listening situation. Usually the speaker adapts the amount of information to the listener by evaluating potential comprehension problems. According to Noblitt (1995), the listening process implies a creation of probable messages out of incomplete acoustic input.

Vocabulary and schemas

Even if the lexicon is unstable by its nature, the understanding of words is considered to be the essence of the comprehension process (Cf. Moss & Gaskell 1999). According to Kelly (1991), limited vocabulary knowledge is the main comprehension obstacle in listening comprehension for learners who master the basic phonologic and lexical code of the foreign language. Word recognition can be thought of in terms of the interaction between acoustic input and lexical knowledge (Dirven and Oakshott 1984). The listener is dependent on the bottom-up process until a suitable solution is found among potential test-takers. Apart from this, word recognition consists of an interaction between acoustic input, vocabulary knowledge and knowledge about the context. The time needed for word identification is dependent on factors such as the frequency of the word, the existence of similar words and the understanding of the meaning of the word and of the syntactic and semantic context. According to some studies a learner needs 1.37 times more phonetic segments than a native in order to recognize multisyllabic words (Dirven and Oakshott 1984).

In their studies based on listening comprehension test items, Freedle and Kostin (1999) have arrived at a result that surprises them: the more multisyllabic words there was in a spoken text, the easier was the test item associated with that text. The researchers have not found an explanation to that phenomenon. My guess is that the multisyllabic words are easier to identify since they exclude ambiguities that will be present in spoken texts with lots of short, unstressed words, colloquialisms and idiomatic expressions, due to the use of phonetically reduced forms (cf. Henrichsen 1984, referred to above).

There seem to be several factors influencing the potential difficulty of a word for a learner. One of the factors is the frequency of the word. Some studies (see e.g. Nissan et al. 1996) indicate that this is the case. Thompson (1995) suggests that texts with frequent words are easier than texts with slang or technical jargon. Anderson & Lynch (1988) have, however, pointed out that words that are frequent for people in general might not be frequent for learners. They stress that the difficulty of a particular word is also related to factors like the context, the listener's knowledge about the treated subject, and similarities between the word and the corresponding word in the learner's L1.

"Lexical bundles" are defined as the most common sequences of words in a text (Biber 1997). Biber used a corpus consisting of five million words in conversations, and established that approximately 40% of all words occur in such lexical bundles. One might conclude that these words are easy to understand because they are frequent, but according to Ur (1984) this is not the case. Many of these collocations are difficult because of the presence of phonetic assimilations and reductions when they are being produced by natives. This assumption is also supported by Anderson and Lynch (1988).

The presence of abstract and modifying words in speech is considered to complicate understanding (see Powers 1985 and Biber 1997). There are different groups of words that modify statements. Some of them reduce the strength of a verb, for instance "almost", "only", "hardly". Others express an uncertainty: "approximately", "more or less", "perhaps". There are also words or expressions that increase the strength of a verb such as "absolutely", "extremely", "completely" and words that express a certainty, such as "surely", "very", "really". These add to the potential difficulty of a spoken text, especially for a learner.

The comprehension of key words or expressions is considered decisive to the relative success or failure a learner experiences with an L2 text. A key word that is repeated or elaborated in speech has, however, different consequences, depending on the learner's level of competence. Repetition is helpful for weaker listeners, whereas elaborations such as paraphrasing, the use of synonyms or appositions are only helpful for learners at higher competence levels (cf. Chaudron 1995; Lynch 1998; Rubin 1994; Thompson 1995).

Having recognized a word, the listener does not just identify it to move on to the next word, but instead he activates a network of words that helps him recognize other related words in the speaker's utterance. In the "mental lexicon" all words are stored with their form and their meaning (Buck 2001:15). The concept of "schema" is also used to refer to the necessary internal, unconscious and individual system we have for organizing our knowledge about a particular theme and for the ability of using inferences (Rost 1994). According to Rumelhart and Ortony (1977) the schema is a basic element in the human information processing system, and a key element in the comprehension process. In our memory, we have several available schemas, both very concrete ones and ones concerning more abstract concepts. It is necessary to realize that a language is also part of a tradition, and we need knowledge and experience of this tradition in order to be able to understand and use the language in an efficient way. An interesting discovery has been made: information that does not fit into the framework of the listener's cognitive schema is not treated at all (Noblitt 1995).

The concept of “script”, the procedure of storing certain episodes in memory, is defined by Schank (1975) as a “causal chain that provides us with knowledge of a frequent situation”. It is with the help of these scripts that we can understand texts, which always consist of a combination of several scripts. By utilizing these scripts as a tool in the top-down strategy, we have the possibility of predicting what will follow in an oral or a written text (Dirven and Oakshott 1985).

Syntax and text organization

The contents of the following paragraphs concern both grammar knowledge and to some extent discourse knowledge as they are categorised in Buck’s framework above. This is explained by the fact that the use of syntax is rarely limited to single, short utterances, but is largely dependent on a larger discourse context. In addition, the target texts that constitute the base for the items analyzed for this study represent pieces of discourse rather than single utterances.

As a trait related to the interactive nature of the listening process, where no single stage can be separated in reality, Dirven and Oakshott (1984) consider the recognition of key words being the probable entrance gate to the syntactic and semantic interpretation of a text. Rost (1994:35) says that the listener, as an aid to the comprehension of utterances, draws on his or her syntactic knowledge in order to be able to divide the text into useful constituents. The listener decides how the words constitute a clause and how clauses make up a sentence. This division helps the listener predict what the speaker intends to say, and to fill in comprehension gaps with words that are found with the help of deductions based on knowledge about syntactic models. This process is, however, too explicit to be accomplished in real-time listening. When a listener divides the utterances into constituents and their relations, he or she does a partial analysis only. Instead of focusing on entire utterances, a listener focuses on the information that is new. Earlier given information, or information that the listener has stored among background knowledge does not need to be analysed.

It has to be kept in mind that the syntax of unprepared speech is different from that of prepared speech. In prepared speech the relations between the ideas or the propositions are expressed mainly through syntax, while unprepared speech partly relies on the context to link ideas together. A speaker sometimes does not mention the referents explicitly, but leaves it to the listener to infer who or what he or she is referring to (Buck 2001: 9-10). The idea units are shorter in speech, while written language is typically denser and syntactically more complicated. The need to mediate as much information as possible in written language usually results in numerous subordinate clauses. In speech the idea units are often linked together with co-ordinate conjunctions (“and”, ”but”, “or”) while written language includes more complicated ways of linking clauses. Among the distinctive features of a spoken text Flowerdew and Miller (2005: 48) list: sentence fragments rather than complete sentences; structured according to tone units rather than in clauses; frequent occurrence of discourse markers at beginning or end; high incidence of questions and imperatives; first and second person pronouns; deixis.

A text can be placed on an oral-literate scale, where oral texts with traits typically associated with everyday conversation constitute one extreme, while literal texts exhibiting more traits of written prose constitute the other (Tannen 1982). Different degrees of "spokenness" or "writtenness" can be seen as a continuum, where spoken texts exhibit linguistic features typical of a spoken (see above) or a written text to a greater or lesser extent (Flowerdew and Miller 2005: 48). It seems likely that the difficulties for a typical classroom-based learner increase when facing a spoken L2 text from either of the two extremes of the continuum. If a spoken text can be placed at the extreme literate end of the scale, the traits caused by for instance complex syntax and abstract concepts make the text more difficult. However, spoken texts containing frequent every-day expressions, *sandhi*-variation and language-specific prosodic traits produced at a relatively fast speech rate, plus some sloppy pronunciation, are no less difficult to understand. How much these features affect comprehension depends on the particular language and the variant concerned, but they certainly play a part in French as a L2 for learners with Finnish or Swedish as their first language. This is due to the differences in the sound landscapes between the languages with the focus in French set at the end of a rhythmic group and in Finnish and Swedish at the beginning of words (see Lhote 1995 and above chapter 1.5.1). To this adds the fact that the stream of speech with the coarticulation, the liaisons and elisions makes the distinguishing of word borders difficult.

A complex but interesting way of covering the dimension of difficulty of the spoken and written registers is proposed by Biber (1995). His model analyses texts along five dimensions: 1) implicated versus informational, 2) narrative versus non-narrative, 3) situation-dependent versus elaborated reference, 4) open versus closed argumentation and 5) concrete versus abstract style. The first mentioned elements of these dimensions are associated with easier texts. Other researchers (Shohamy & Imbar 1991; Brown et al. 1985) have also established that narrative texts or interviews are easier to follow than texts that are more informative, such as speeches. According to Shohamy and Imbar (1991) different types of texts are associated with different degrees of difficulty. It seems, however, that it is not the type of text in itself that influences the difficulty, but the syntactic, discursive and pragmatic features within a certain genre. Freedle and Kostin (1991) have obtained similar results in their studies: test items based on the comprehension of concrete texts containing relatively few abstract concepts turn out to be easier for the L2 learners.

A look at morphologic and syntactic modifications aimed at facilitating understanding can give indications on where the factors of difficulty in a text are situated. Several studies seem to point to the fact that there are various integrated variables affecting understanding: the listener's competence level and the amount of background information possessed, the type of modification undertaken as well as the type of text passage. In this connection it is essential to consider how the syntactic complexity of a text can be measured. Many studies have been conducted on isolated syntactic problems. Anderson and Lynch (1988) and Bygate (1987) suggest that co-ordinating structures are easier to understand than subordinating structures. Cervantes and Gainer (1992) have used the amount of s-nodes (subject-verb constructions) per unit (clause or sentence) as a measure of complexity. In their study, they were not, however, able to find any significant differences between two groups of learners, one of which listened to a text with simplified syntax while the other was presented with a text containing more complex syntax.

Freedle and Kostin (1999), basing their conclusions on studies with English as a L2, have established that the presence of negations in the text or in the alternatives is related to more difficult MC items. They have also concluded that pronouns or nouns and references that range over multiple sentences cause comprehension difficulties. Hansen & Jensen (1994) on their hand have shown that several different syntactic structures are bound to increase the complexity of a spoken text: nominalizations, indirect questions, relative complementary or restrictive clauses, as well as adverbial or prepositional clauses.

There appears to be various attempts to describe syntactic complexity. However, since no one variable has been singled out to determine this complexity alone, it is reasonable to think about the phenomenon in terms of an accumulation of variables functioning together.

1.5.2 Discourse knowledge in listening – processes and problems

Having discussed how syntactic features as part of a larger discourse context can affect the listening comprehension in different ways I will now present some further aspects of the treatment of larger chunks of spoken discourse by looking at the effects that redundancy, elaboration and coherence have on comprehension.

It seems reasonable to look at listening texts in larger contexts, that is, from the point of view of their organization. As Buck (2001:113) points out, listening test tasks should require listeners to process more than just short utterances. Including longer texts, he says, tends to engage aspects of pragmatic knowledge and strategic competence, which may otherwise be ignored ingredients when listening frameworks for learning and assessment are too limited.

Even if it is usually supposed that listening comprehension is difficult for foreign language learners simply because aspects of the language are difficult, there is a further component to be considered: the content of what is said (Brown 1995: 59-69). Based on the concept of cognitive load, “commonsense parameters”, which have been shown to make the language in different text types (narrative, descriptive, instructional, and argumentative) more or less difficult to understand, can be given. Brown mentions the following principles:

1) It is easier to understand any text that involves fewer rather than more individuals and objects.
2) It is easier to understand any text involving individuals and objects that are clearly distinct from one another.
3) It is easier to understand texts involving simple spatial relations.

4) It is easier to understand texts where the order of telling matches the order of events.
5) It is easier to understand a text if relatively few familiar inferences are necessary to relate each sentence to the preceding text.
6) It is easier to understand a text if the information in the text is self-consistent and fits readily with information you already have.

However, as Oxford et al. (2004: 12) points out, not every cognitively complex task is viewed as difficult:

> Whether or not a particular student actually perceives a given, cognitively complex task to be difficult and challenging depends considerably on the student's familiarity with doing the kind of cognitive operations required. If a student has had lots of practice with a complex task, then doing another task of a similar kind might be straightforward because of familiarity.

Redundancy is one important way in which the elements of language relate to each other (Buck 2001:67). Speech is redundant on many different levels: on the acoustic, phonological, morphological and syntactic levels, and further increased by the context, the co-text, the topic and the situational context. As speakers naturally modify their speech depending on the situation, for example speaking more slowly and with clearer enunciation when speaking to someone who has less background knowledge, it is interesting to consider how redundancy variables affect listening comprehension. One of the researchers who have investigated these phenomena is Chaudron (1983) who found out that redundancy in the form of a repetition of a noun facilitates understanding and memorization. The degree of redundancy in a text has consequences for the ability of the listener to understand the given information. With plenty of redundancy there is not as much new information to treat. A text with very limited redundancy is often difficult to understand, especially if the theme of the text is unfamiliar to the listener.

There are other ways of modifying speech. According to Kelch (1985), it is only in combination with a slow pace that lexical modifications (a replacement of uncommon words with synonyms) and syntactic modifications (replacement of utterances with paraphrases) support understanding. Chiang and Dunkel (1992) have found that only learners on an advanced level can profit from modifications such as redundancy of information and elaboration (repetition of determiners and nouns, use of paraphrasing and synonyms) but not learners on lower levels. Thus two variables seem to determine if redundancy facilitates comprehension: the listener's level of competence and the type of redundancy. For learners on an advanced level a high degree of redundancy constitutes a facilitating factor, whereas the situation is the opposite for beginners. If a text includes a large amount of background detail, its redundancy is more problematic for less advanced learners. An explanation is that beginners faced with texts that are heavy with detail will have to treat a greater amount of information with limited linguistic resources (Chaudron 1983; Glisan 1985; Pica et al. 1987; Lynch 1988; Derwing 1989; Chiang and Dunkel 1992).

In one of her studies, Derwing (1996) investigated the effects of elaborations produced by native speakers interacting with learners. One of the conclusions was that irrelevant details made the primary message hard to follow. Another conclusion concerns the effects of elaborations: if, as a consequence of the elaboration, the entire structure of the speech act is altered – affecting text coherence and text cohesion – the effect will by no means be facilitating for the L2 listener.

The effects of linguistic markers in a prepared speech based on a written document, i.e. a type of text often lacking these markers, were studied by Chaudron & Richards (1986). They wanted to prove that certain markers that guarantee the cohesion (micromarkers or links between sentences or clauses) and the textual coherence (macromarkers or links between parts of the discourse) facilitated the comprehension of this type of speech. It appears to be decisive for the relative difficulty of a text what type of cohesive marker is used. The listener automatically tries to find coherence between several utterances – links between the ideas in a conversation. Coherence is found to the extent to which the listener is able to use his/her background knowledge and expectations, combined with the speaker's use of verbal cohesive markers. Some cohesive devices that are efficient in a written text, do not, however, facilitate listening to a spoken text. The cohesive markers may actually create long sentences with several subordinate clauses, which are typical for texts situated at the literate extreme of the oral-literate scale. In Freedle and Kostin's (1999) research there were three rhetoric text organizers that influenced item difficulty: the use of a list resulted in easier items, while the construction "problem-solution" and "comparison" resulted in more difficult items.

1.5.3 Pragmatic and sociolinguistic knowledge

We interpret language in a context, and the interaction between what is said and the context in which it is said is an important aspect of comprehension (Buck 2001:22). As Buck's listening framework suggests, by adding the pragmatic stage to the listening construct, we take one further step towards a listening ability similar to that of a native speaker. For construct definition purposes Buck (2001:105) defines this aspect as "understanding inferred meanings", including the two previous stages (grammar and discourse).

With the addition of the sociolinguistic dimension, which means that the listener understands language elements as they appear in a particular social and cultural situation, we have reached the communicative listening ability that can be said to be the ultimate goal for a language learner who seeks to obtain a listening comprehension competence with maximum coverage. It is, of course, only on more advanced proficiency levels that this ability can be fully reached.

1.6 Listening comprehension strategies

In the listening framework proposed by Buck (2001: 104) and quoted above, the language competence part is completed with a strategic part, which includes the ability to use the language competence, consisting of *"executive processes that fulfil the management function in listening"*. It is thought of as a "*mediator between the external situational context and the internal knowledge*" (Bachman 1996). The question arises of how to define strategies as contrasted to processes and the subcategory of cognitive strategies as contrasted to the subcategory of metacognitive strategies. Moreover, in the context of the current study, it appears essential to consider the differences between comprehension strategies proper as separated from test-taking strategies.

1.6.1 Strategies vs. processes

The meaning of strategy can be traced back to Greek military terminology, and the significance of a plan to win a war (Van Dijk & Kintch 1983: 62; Oxford 2003: 274). In broad modern usage, across a wide range of domains, the term assumes conscious control, intention and goal-directedness – "*reaching a goal in an optimal way – quickly, efficiently and with low cost*". Van Dijk and Kintch (1983: 69) further explain that a strategy will in general be the result of a mental process involving much information. As soon as this mental process is consciously controlled such that each mental step yields the information necessary for the next mental step, we may also speak of mental strategies. They point out that comprehension normally is an automatized activity, not monitored at each step by the language user (ibid: 70). Nevertheless, they explain why strategies, together with language competence, are necessary ingredients for natural language processing. They list the following processing features:

1. Language users have limited memory, especially short-term memory capacity.
2. Language users cannot process many different kinds of information at the same time.
3. Understanding language is linear, whereas most of the structures the language rules pertain to are hierarchical.
4. Understanding requires not only linguistic or grammatical information, but other information as well, about the context, episodic memories, knowledge of the world etc.

Strategies are thus necessary to allow a language user to accomplish the task of understanding linearly, at several levels, simultaneously taking into account different kinds of information, and with limited knowledge (ibid: 72-73). This holds especially true in a foreign language situation, where even fewer stages of the comprehension process may be automatic.

There is some variation in the definitions of processes and strategies. Purpura (1999) follows Bialystoc's (1990) distinctions in delineating between processes and strategies: accordingly, processes would be conscious or unconscious mental steps taken to carry out a cognitive activity, while strategies are techniques or tactics used to carry out these processes.

The question of the degree of consciousness can be approached in a slightly different way. Phakiti (2003) follows the ideas of Cohen (1998a) and Ellis (1994) by describing strategies as conscious acts that can be accessible for description, whereas "common" processes can be unconscious and automatic. However, she points out that consciousness about strategy use may vary between individuals. Some strategies may thus become processes for some individuals, remaining strategies for others. Potentially, therefore, in introspective or retrospective reporting on an individual's strategy use, the non-mentioning of a particular strategy may be due to either an inability to make use of a strategy (more probable among low-ability learners), or an automatizing of that strategy to an unconscious process (more probable among high-ability learners). According to Oxford (2004:1), following Cohen (1998a) and Goh (2002), most theorists and researchers agree that some degree of consciousness or awareness is essential in strategy use.

Phakiti (2003; cf. Cohen 1998a) points out the fuzziness in the definition and the subcategorizing of the notion of strategies. As a basic principle, learner strategies can be divided into two types: learning strategies and use strategies. Use strategies, essential for the present context of testing listening comprehension, are associated with a particular situation and are purposefully employed strategies to enhance performance – whether in a communicative or a test-taking situation. Cohen (1998, according to Oxford 2003: 275) describes four types of language use strategies: strategies for retrieving information about the L2 already stored in memory, strategies for rehearsing L2 structures, cover strategies to help the learner avoid looking stupid or unprepared, and strategies for communicating in the L2 despite knowledge gaps.

Drawing on the definitions and conclusions of the researchers mentioned above, the main characteristics related to the concept of strategy thus seem to be: consciousness, control and goal-directedness. In the following discussion, this view on how to separate the two concepts of processes and strategies will be maintained. I am aware of the impossibility of drawing an exact boundary between the two, as one may, in some circumstances, face a difference of degree rather than one of category.

1.6.2 Cognitive vs. metacognitive strategies

There are strategies of two main kinds. Bachman & Palmer (1996) define cognitive strategies as such mental activities that are related to comprehending and storing input in working memory or long-term memory for later retrieval. The metacognitive strategies are the mental activities that perform an executive function in the management of cognitive strategies. Thereby whereas a cognitive strategy is used to reach a specific goal, a metacognitive strategy is employed to check whether the goal was actually attained (Stemmer 1991: 235).

Added to these two, socio-affective strategies are related to managing emotions and to learning with others. As examples of these strategies Flowerdew and Miller (2005: 78-79) mention questioning for clarification, cooperation, lowering anxiety, self-encouragement and taking emotional temperature.

In a test-taking situation, metacognitive strategies are the test-takers' deliberate mental behaviors for directing and controlling their cognitive strategy processing for successful performance (Phakiti 2003: 30). Some of the strategies treated above are pertinent for a learner in a listening situation, and these are discussed in the following section.

1.6.3 Listening comprehension strategies

O'Malley & Chamot (1990), Young (1997) and Goh (2002) have created inventories of listening strategies or tactics, partly parallel, partly drawing on different principles for their classification (see Table 4). The objective of a study conducted by O'Malley & Chamot (1990) was to find out if the strategies students used paralleled the comprehension phases of *perceptual processing, parsing* and *utilization*. Different strategies were found and related to each of the three processing phases. Young (1997), inspired by the works of O'Malley & Chamot (1990), Oxford (1990), Rost & Ross (1991) and Vandergrift (1992), presents a comprehensive Second Language Listening Comprehension Strategy Inventory, illustrating the many subskills involved in the process. Goh (2002) speaks about cognitive and metacognitive tactics related to learner listening, and her inventory of tactics is more or less identical to the general strategies – cognitive and metacognitive – described by Purpura (1999), Buck (2001) and Vandergrift (2003).

The following table (Table 4) represents an attempt to show the parallelism and the slightly differing approaches across these three inventories. They reflect the idea of the cognitive processes and strategies involved in a listening comprehension situation as being active and goal-directed. In the inventories by Young (1997) and Goh (2002), the difference to O'Malley and Chamot (1990) is in the inclusion of the strategies preparing for the listening activity to come. Young (1997) also differs from the other two in that the last six strategies in Table 4 imply the presence of an "external resource", that is, a textbook or the interlocutor. In most test-taking situations this would not be possible.

TABLE 4 Listening comprehension strategies as classified by Young (1997), Goh (2002) and O'Malley & Chamot (1990)

Young (1997)			**Goh (2002)**	**O'Malley & Chamot (1990)**
			Pre-listening preparation (Metacognitive)	
Planning	Developing an awareness or an action plan of what needs to be done to accomplish a listening task, or making predictions while listening		Predicting (Cognitive)	
		PERCEPTUAL PROCESSING	Selective attention (MetaC)	Selective attention
Self-monitoring	Checking, verifying, or correcting one's listening comprehension while performing a task		Comprehension monitoring (MetaC)	Self-monitoring
			Real-time assessment of input (MetaC)	
			Fixation (C)	
Grouping	Classifying information such as words or concepts according to their meaning or according to the listener's own organization	**PARSING**		Grouping (listening for larger chunks)
Inferencing	Using acoustic, vocal, or lexical information within the text to guess the meaning of unfamiliar language items, or to fill in missing information		Inferencing (C) Contextualisation(C)	Inferencing from context
Repetition	Repeating a word or phrase in the target language mentally or orally		Reconstruction (C)	
Imagery	Using mental or visual images to represent information		Visualisation	
Elaboration	Activating prior knowledge outside the text or conversational context to construct meaning or to fill in missing information	**UTILIZATION**	Elaboration (C)	Elaboration from world knowledge, personal experiences, or self-questioning
Self-evaluation	Checking the concentration of oneself or assessing one's strategy use		Comprehension evaluation (MetaC)	
Problem Identification	Pointing out the central point needed to be resolved in a task, or identifying an aspect of the task that hinders its successful completion			
Translation	Expressing target language words in the listener's first		Translation (C)	

	language in the listening task	
Transfer	Using knowledge of one language to facilitate listening in another	
Deduction	Reaching a conclusion about the target language because of other information the listener thinks to be true	
Summarization	Making a mental or oral summary of the information presented in a listening task	
Note Taking	Writing down key words and concepts while listening	
Resourcing	Using available references about the target language, including textbooks or the previous tasks	
Reprise	Showing the "speakers" that they did not get the message across	
Feedback	Giving comments about the aural text	
Hypothesis Testing	Asking specific questions about facts in the text to verify one's schematic representation of the text	
Uptaking	Using kinesics and paralinguistics to signal the "interlocutor" to go on	
Clarifying	Asking for explanation, verification, rephrasing, or examples about the language and/or task, or posing questions to the self	

In addition to these inventories, Nakatari (2006) has developed an Oral Communication Strategy Inventory including eight categories of strategies for coping with speaking problems and seven categories of strategies for coping with listening problems during communication. While some of the listening strategies imply the element of interaction, others are general by nature. There are “fluency maintaining” strategies like paying attention to the speaker’s rhythm and intonation or pronunciation and “scanning” strategies like paying attention to the subject and verb of the sentence or to the interrogative, to the first part of the sentence to guess the speaker’s intention or trying to catch the speaker’s main point. Among the “getting the gist”- strategies, Nakatari mentions ignoring the problem of not understanding every single detail, the anticipation of what the speaker is going to say based on the context by guessing the speaker’s intention on the basis of what has been said so far. The category of “less active listener”-strategies, which can often be counterproductive, includes the attempt to translate little by little into native language and to focus on familiar expressions only. By means of the “word-oriented”-strategies the listener pays attention to the words that the speaker slows down or emphasizes, guessing the speaker's intention by picking up familiar words, trying to catch every word that the speaker uses, or paying attention to the first word to judge whether it is an interrogative sentence or not (Nakatari 2006: 167-168).

Rost (2002) as well as Young (1997) have empirically or through a meta-analysis of previous studies listed the most frequent sequences of uses of strategies, mainly employed in this way by more successful or more advanced listeners. Rost (2002: 155) refers to studies by Vandergrift (1996; 1998; 1999) and Rost (1999) stating that there is broad agreement on the kinds of strategies that are associated with successful listening. The following six strategies are mentioned:

1) Predicting information or ideas prior to listening.
2) Drawing inferences about complete information based on incomplete or inadequate information.
3) Monitoring one's own performance while listening, including assessing areas of uncertainty.
4) Formulating clarifying questions about what information is needed to make a fuller interpretation.
5) Providing a personal, relevant response to the information or ideas presented.
6) Evaluating, checking how well one has understood, and whether an initial problem posed has been solved.

Through the analysis of data collected by think-aloud procedures of 18 advanced Chinese undergraduate ESL students listening to audio texts, Young (1997) was able to establish that listeners who used relatively more types of strategies – generally higher-ability listeners – used the following five strategies in a serial order/sequence in the following way:

1) Through inferencing from contextual or acoustic clues, the students guessed the theme or topic of the text.
2) Through elaboration they activated their background knowledge of the topic.
3) They then used summarization to reinforce their own interpretation of the text.
4) The metacognitive strategies of self-monitoring or self-evaluation were used to control their comprehension and to evaluate their strategy use.
5) Feedback was given in interacting with the text.

This series was employed in the same serial order throughout the tasks, a pattern that also occurred with students who used fewer strategies.

Although our understanding of the complex processes involved in listening comprehension strategies may be limited, Vandergrift (2003) maintains that the research literature on such strategies points to some useful findings for both content and methodology: a) metacognitive strategies such as selective attention and comprehension monitoring are reported more frequently by more skilled listeners; b) cognitive strategies such as elaboration and inferencing are used almost equally by all listeners but appear to be used in more effective combinations by more skilled listeners; c) more skilled listeners appear to be more flexible in strategy use, employing strategies in effective combinations; d) the three phases of the listening process (perceptual processing, parsing, utilization) can be identified in listener think-aloud protocols, as well as strategies associated with each phase; e) a think-aloud procedure appears to be a productive methodology for studying on-line strategy use; and f) a qualitative analysis of protocols, in addition to a quantitative analysis, appears to provide greater insight into the differences between more skilled and less skilled listeners (Vandergrift 2003: 470-71).

Inferencing is a key process (or strategy, depending on how one classifies this cognitive activity), mentioned by all theorists and researchers as a necessary ingredient for communication and thus essential for successful listening comprehension. Psychologists see the process of making inferences – of ”going beyond” the input - as a general characteristic of humans, needed to make sense of everything we see or hear (Greene 1986: 24). Haastrup (in Faerch & Kasper 1987: 197) describes inferencing as making informed guesses as to the meaning of (part of) an utterance in the light of all available linguistic cues in combination with the listener's general knowledge of the world, his or her awareness of the situation and relevant linguistic knowledge.

Based on Vandergrift's (1997) taxonomy, Flowerdew and Miller (2005: 75) have listed different listening strategies in the classroom. They place inferencing among the cognitive strategies, and divide it into five types on the basis of inference support. Linguistic inferencing is defined as guessing the meaning of unknown words by linking them to known words. Voice inferencing is also helpful, as well as paralinguistic, or kinetic, inferencing. Extralinguistic inferencing is linked to the requirements of a task, and inferencing between parts implies making use of certain words not directly related to the task to get more information about the text.

One of the factors enabling us to listen in an efficient way is our capacity for prediction and expectations on what the speaker is going to say. Like a listening strategy, the process works according to the principle of “minimal attachment”: a constituent is linked to the preceding one with as simple a structure as possible. The human ability to fill in information gaps is called “analysis by synthesis” in psychological terms (Noblitt 1995). The listener also infers retroactively in order to fill in the gaps in preceding utterances. A parallel treatment of the text is activated: a possible interpretation of the words is carried out simultaneously with an interpretation of the syntax. The listener's knowledge of the world influences understanding in many ways, for example through the general context of the text that limits the amount of possible interpretations, and through knowledge of specific facts and the established relations between things (Buck 2001: 18). Thus one part of the process of inference by the listener is achieved through conventional procedures involving language use, and another part is achieved through problem-solving procedures involving logic and real-world knowledge (Rost 2002: 64).

Logic reasoning is needed to understand what is going on around us, and it is also needed for the comprehension process. In fact, a great deal of the meaning of an utterance is not explicitly presented. The context and the implicit parts are important communication ingredients, which by logic reasoning complete the incomplete explicit information. All users of language must rely on the assumption that as much as 90 % of what might be stated need not to be stated but can be inferred by listeners (Brown 1995: 68).

1.7 The effect of the learner's/listener's level of skill and personal characteristics on the use of strategies

For the purpose of the current study it is relevant to consider what is stated about the possible differences in the processing of spoken input as a function of the skilfulness and the personality factors of the learner. If we summarize the findings on the typical strategy uses of skilled listeners on the one hand, and less skilled listeners on the other hand, there seems to be some agreement on the following characteristics (see Table 5):

TABLE 5 Hypothetical description of a skilled listener vs. a less skilled listener

	Skilled listeners:	**Less skilled listeners:**
Processes	Many basic comprehension processes automatized; fruitful interaction of bottom-up and top-down processes	A great amount of bottom-up processing needed, also at the perception level; sometimes over-reliance on top down processes (letting beliefs override actual text contents); potential problems at all processing stages
Strategies	Generally more varied strategies, applied more frequently, more flexibly and adequately: -Prediction of theme of input before listening; -Background knowledge activated; -Selective attention to text or speaker's prosody; -Not getting distracted by unknown words; -Inferencing in case of incomplete information; -Monitoring performance, evaluation of strategy use; -Relevant response/feedback to presented text; -Evaluation of the success of the task	Fewer types of strategies, often compensating by nature, employed less frequently: -Attempt to translate little by little into native language -Focus on familiar expressions only -Get stuck on comprehension problems/obstacles.

The development of the listening skill from a beginner learning a language to an advanced listener forms an interesting issue. Rost (1994: 137-138) presents the process by describing listeners on three levels. As he puts it, evidently it is partly a question of quantity, of larger repertoires of vocabulary and syntactic constructions. There are certainly other dimensions to consider. Another description of the stages in the development of the skill of understanding speech in a foreign language is found in the *CEFR* (2001). This description has a very practical point of departure, and is meant to provide a framework both for self-assessment – for an L2 learner or speaker to know where he or she stands in comparison with other speakers of that language or with his or her ability in another language, or from a pedagogical point of view to know what potential developmental steps there are to reach – and as a framework to which language skills and assessment of these skills in different languages throughout Europe (and beyond) can be related.

As is described in the *CEFR* by levels ranging from A1 to C2[15], the development of the skill of listening comprehension proceeds through several dimensions. The basic assumption is that language learners move from immediacy and concreteness to distance and abstractness. This is operationalized through the type, amount and complexity of the language that can be understood, the topics that can be treated, to the type of information contents and type of media that can be handled. According to the description of the characteristic "behavior" and skills, this also concerns the way reception strategies can be made use of. From familiar words and concrete concepts, personal surroundings and interests, slow and deliberate speech on A1, we move towards the ability to understand the main points of simple messages and to make use of the overall meaning to understand words on A2, to topics within the job-related domain and everyday matters, straightforward factual information and sentence meaning deduction on B1. On the level B2 extended speech and more complex lines of argument on abstract and not directly familiar topics can also be handled, within the social, academic and vocational domains. To this is added the ability to use a variety of reception strategies and to distinguish the speaker's viewpoints, mood and attitudes. At the level C1 even idiomatic expressions and colloquialisms can be managed, finer points of detail can be sorted out, and attitudes and relationships between native speakers in discussions can be understood. Contextual cues and anticipation are used as strategies to catch the meaning of the most complex speech samples. At the highest level, C2, near-native proficiency is attained according to the descriptors of the levels. The target level for the Finnish Matriculation Examination of French as a L2 is at A2.

Individual listener characteristics are bound to influence the listening process and its outcome in various ways. Among the characteristics that may be decisive, besides the cognitively determined factors (language competence level, background knowledge, short-term memory capacity) Cornaire (1998) mentions

[15] See Appendix V

the following learner characteristics: degree of attention, affect (in terms of attitudes, beliefs, emotions, self-confidence), age and sex.

Since learners are oriented toward listening activities through their learning styles, their preferred learning styles affect the listening process. Flowerdew and Miller (2005:62) have borrowed from Kyriacou et al. (1996) the dimensions and descriptions of potential differences in learning styles. The dimensions range from covering the attitude towards learning (deep, surface or strategic learners), the ways knowledge is built up (holistic vs. serialist learners), the degree of creativity (divergers vs. convergers), the preference for concreteness or abstractness, the attitude towards problem solving (reflective or active learners) to the social aspect (solitary or social learners).

Flowerdew & Miller (2005:64) point out that these learning styles represent the ends of scales, and that the placement of a learner on a scale depends not only on personality, but also on more transient factors like the task type, the time of day, level of interest and motivation.

The assumption in the context of L2 listening comprehension and the assessment of this ability is that these differences in personality are likely to affect the ways in which individual learners and test-takers tackle a task. This in turn plays a role in the appropriate selection of strategies in a specific listening situation. According to research by Rantanen (2003: 102, 187) a test-taker's personality and confidence also affect his or her tendency to guess the response to a MC item.

1.8 Test-taking strategies in listening comprehension contexts

There are some essential differences between general listening strategies and those employed in the context of a particular test situation. According to Cohen's (1998b: 92) definition test-taking strategies are language use strategies applied to tasks in language tests but they also include test-wiseness strategies (Cohen 1998a: 219). All four types of language use strategies – retrieval, rehearsal, cover and communication strategies - are used in test-taking. Test-wiseness strategies are not necessarily determined by proficiency in the language being assessed, but may be dependent on the knowledge of how to take tests (Cohen 1998a: 219). Among these strategies there are cases like opting out of the language task at hand, for instance when matching surface information in a text with information in an option, making use of material from a previous item that "gives away" the answer to a subsequent one, or taking shortcuts to arrive at answers, for instance when not reading an entire text as instructed, but only limited passages to answer questions. Moreover, a MC option may be selected because of its differing form, its placement or its order. In these cases, test-takers use test-wiseness in order to avoid reliance on language knowledge, especially if the linguistic skill is felt to be inadequate for the task.

The employment of test-taking strategies will thus vary with test format. Cohen (1998b: 93) points out that the use of a limited number of strategies in a response to an item may nevertheless indicate genuine control over the item, assuming that these strategies are well-chosen and are used efficiently.

Phakiti (2003) describes how the metacognitive strategies of planning and monitoring are used in a test-taking context. For her, metacognitive planning strategies used by test-takers are those directed at the regulations of the course of their own thinking, and which help to:

- allocate resources to the current task;
- determine the order of steps to be taken to complete the task; and
- set the intensity or the speed at which one should work on the task.

Monitoring strategies are the test-takers' deliberate actions to check, monitor and evaluate their thinking and performance. Monitoring strategies help to:

- identify the task on which one is currently working;
- check on the current progress of the work;
- evaluate the progress, and
- predict what the outcome of the progress will be (Phakiti 2003:30).

1.9 Listening comprehension tasks

The situation or the task at hand determines how the listening processes are activated and how the strategies are possibly used. The listening framework summarized by Buck (2001) illustrates a competence-based listening construct. Buck (2001) also suggests the possibility of describing the listening construct on the basis of the tasks that the listener is supposed to be able to perform. This approach has its advantages as well as its disadvantages. From the point of view of the selection of convenient tasks for measuring the skill, it would appear easy to select samples of real-world, target-language use tasks that serve as a basis for describing the skill. What is more, it would have practical implications for the ability to use the skill in real-life situations: if a test-taker manages a task in the test, (s)he will probably manage that task in a real-life situation just as successfully.

However, the question arises how we are to limit the tasks or the language use situations that are to be included in a construct. The tasks are naturally closely linked to authentic and practical communicative situations. There may be clear restrictions in some cases, for example within the military context, where the goal may be to assess the capability of understanding orders in the target language. It is fairly easy to envision the limited vocabulary and syntax needed in such situations. Most communicative situations, however, can be highly unpredictable, include interlocutors, and be combined with tasks related to skills that seem to be outside the listening construct proper. In these cases we face the difficulty of limiting, in a meaningful way, the description of the construct to be measured (cf. Messick 1994).

Takala (1998) draws an overview comparing the relative advantages and disadvantages between authentic, or alternative, or performance assessment[16] and traditional assessment. The table below replicates Takala's summary (Table 6):

TABLE 6 Advantages and disadvantages of authentic/alternative and performance assessment (Takala 1998)

Feature	**Aim/goal/intention**	**Potential strengths**	**Potential criticisms**
Authentic (alternative, performance) assessment	Assessment must reflect a "modern" view of learning and the natural uses and contents of knowledge	*Important and valuable goals are assessed *Assessment is in line with the curriculum and even supports its attainment *Assessment is felt to be meaningful and motivating *Assessment reflects a person's strengths and may bolster self-image	*Authenticity is not an unequivocal concept and thus does not have unequivocal criteria either *Alleged benefits of authentic assessment lack strong, solid evidential basis
Traditional (multiple-choice-based) assessment	Assessment should, above all, be reliable and commensurate – the context of use is secondary	*Subjectivity is under control *Reliability is generally good *The domain to be assessed is covered well *Assessment is cost-effective	*Validity can be a problem *Washback effect on teaching may be undesirable *Assessment may focus too much on memorization, and larger knowledge structures may be neglected
Degree of directness of assessment (testing):	Assessment must reflect its target as closely as possible; the effect of target-irrelevant factors should be minimized		All assessment is indirect and always requires interpretation
More direct		*Face validity of assessment is good *Interpretation of results is more clearcut (low-inference)	*Scoring requires 'subjective' judgement (methods variance)

[16] See also for example Bachman 2000:11-15 and Messick (1994).

More indirect		*Probably a better control of assessment target *More objective scoring	*Face validity weaker *Interpretation of results less clear-cut (high-inference)
Assessment based on tasks (task-driven)	Enhancing the ′pragmatic′ aspect of validity	*Assessment is credible since authentic tasks allow, and require, the use of all important skills and knowledge necessary for a good performance	*It is not easy to define tasks in an unambiguous manner *It is not clear how generalizable information is obtained by task-based assessment
Assessment based on the cognitive basis of knowledge and skills (construct-driven)	Enhancing the ′conceptual′ aspect of validity	*Assessment is generalizable, since it is known what the tasks are based on	*Interpretation is not as straightforward as in task-based assessment
Assessment based on a very open situation	Enhancing "real-life" linkage	*Assessment corresponds well to "real life" where the situations are often "open" and a person has to decide for him/herself what it is all about	*Openness may baffle and lower performance for some individuals *Openness is relative – even partly structured situations may be close to "real life"
Assessment based on a highly structured situation	Enhancing reliability and control of error	*Assessment situation is well under control: diagnostic information is obtained at desired level of accuracy ("grain") *Restricted assessment situation creates a sense of security	*Assessment is artificial and does not provide an adequate picture of proficiency *Structured situation may be felt to be too restrictive, which may lower motivation

Buck (2001) recommends thinking of the tasks in terms of the characteristics of the context in which the listening takes place. The task characteristics can be compared with real-world tasks. The important issues include the notion of authenticity – the extent to which the test tasks are perceived to share the characteristics of the target-language tasks or to engage the same abilities as the target language tasks (Buck 2001: 106). The idea is thus to replicate real-world tasks as closely as possible. This is, however, often difficult, as a test-taking situation typically distorts the "natural circumstances" of a real-world task. Bachman & Palmer (1996:25) talk about interactiveness – the extent and type of involvement of the test-taker's [or any language user's] individual characteristics in accomplishing a test task [or any language use task]. Thus the interactiveness of a given language task can be characterized in terms of the ways in which the test taker's areas of language knowledge, metacognitive strategies, topical knowledge and affective schemata are engaged by the test task.

In (reading and) listening tasks the type and difficulty level of a task and the amount of advance planning would logically affect

1) the fluency or ease with which the input is comprehended
2) the accuracy of the understanding, as well as
3) the complexity and richness of mental frameworks (schemata) by which the read or heard material becomes stored in memory, ready for later use (Oxford 2004:13).

In their discussion of performance-based assessment Norris et al. (1998) note that sampling of L2 performance assessment tasks should come out of an understanding of the processing attributes inherent in needs analysis- selected real-world tasks. I will discuss this issue further below in the chapter on the test format (chapter 2).

1.10 The implications for the current study of the description of listening processes and strategies

A combination of the theoretical frameworks and research results referred to above yields a processing model – in the restricted sense of one-way processing - involving at least the following elements or dimensions related to the listening comprehension processes, presented in Figure 2 as a schematic graphic figure:

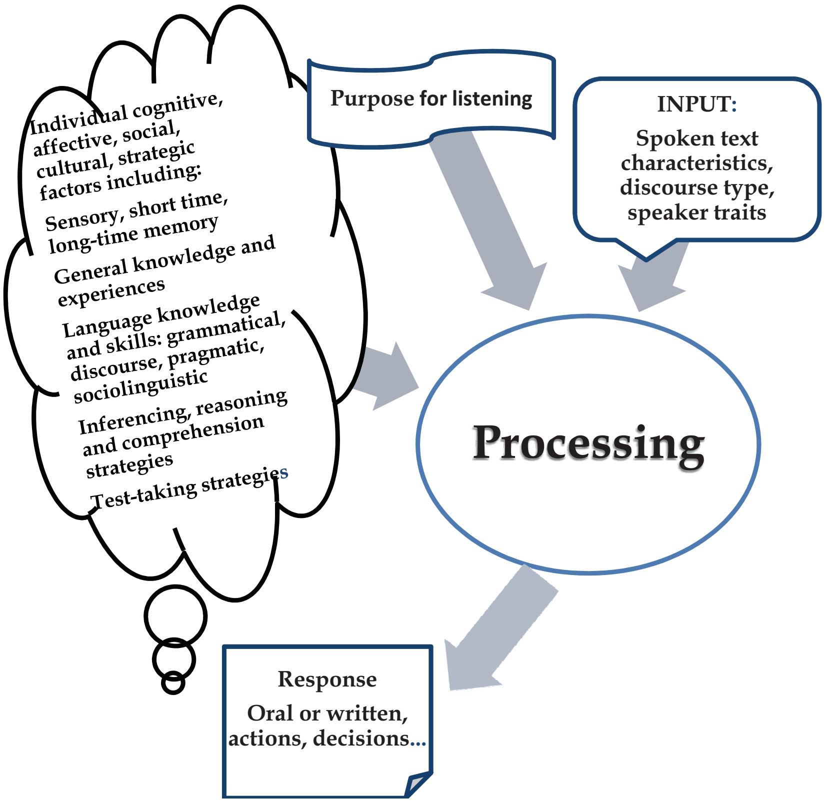

Individual cognitive, affective, social, cultural, strategic factors including:
Sensory, short time, long-time memory
General knowledge and experiences
Language knowledge and skills: grammatical, discourse, pragmatic, sociolinguistic
Inferencing, reasoning and comprehension strategies
Test-taking strategies
Purpose for listening
INPUT:
Spoken text characteristics, discourse type, speaker traits
Processing
Response
Oral or written, actions, decisions…

FIGURE 2 Simplified processing model

It should be pointed out that an interactional model of processing would have further dimensions; here my attempt is to take into account the elements treated mainly from a cognitive point of view in the framework presented in the preceding chapters.

The essential discussion in this chapter concerns the question of what is to be included in the construct of listening comprehension. A number of generalities as well as a number of specificities need to be taken into account. For first-language use, both language competence and strategic competence are relatively well developed, and linked together in a long-term stable relationship (Buck 2001:103-4). In second-language testing, it would therefore seem reasonable to put the emphasis on testing language competence rather than strategic competence, since differences in performance between individual listeners will generally be due to differences in linguistic competence. A learner's cognitive abilities will, on the contrary, be relatively stable and ready developed. However, it seems obvious that a second-language learning situation and, to an even higher degree, a test-taking situation, are *per se* strategic, and it would be difficult to exclude the effect of strategic competence on the performance of an individual test-taker. The way the L2 has been taught and learnt enters as an important factor: has the teaching focused on more traditional skills of understanding details and getting the gist of spoken texts or have the classes included activities raising the learners' awareness of listening strategies as well?

A few other key issues that serve as bases for the further discussions need to be pointed out. First of all, it is clear that the general process of foreign language listening, making sense of spoken input presented in a foreign language, is anything but a simple, linear straightforward activity. Many subskills are involved, many intra-individual aspects are decisive for the nature and the outcome of the activity which is characterized by an interaction and integration of bottom-up and top-down processes on all stages, from auditive perception and discourse comprehension to pragmatic interpretation.

Secondly, the task, the context and the situation are crucial factors for the processing and the outcome thereof. This holds true whether the task is set in a real-life target language use (TLU) situation, in a classroom setting or in the context of a test. Thirdly, related to the situational, contextual or task-based aspect of the listening process, the presence of a task, a goal, or a target for the activity makes it more or less strategic. Language use is a strategic, cognitive activity. Taking a test and solving test tasks naturally involves both strategies that are closely linked to the language processing, and strategies that are linked mainly to solving a certain type of task. Test-wiseness is an example of the latter types of strategies

Fourthly, taking into account the fact that the listening process is complicated, internal and hidden, but that test constructors or users still want to evaluate this ability, and do it in a valid way, one question is how we can find out more about the nature of the processes. We want to make sure that we are targeting the intended construct. As direct observation is impossible, I will investigate the present methodology of introspection (treated below in chapters 3 and 6) as a possible means of finding out more about the listening processes and the variables affecting them. In the following chapter, I will treat the nature of general test tasks and those used for assessing listening comprehension ability. The test tasks and items provide a more or less authentic context or purpose for the listening activity.

2 THE CONSTRUCT, THE TEST FORMAT AND THE VALIDATION PROCESS

In the previous chapter I described the general listening process and strategies related to that process and to more test-specific contexts, shedding light on the construct behind foreign language listening comprehension. If we want to assess that ability, there are, nevertheless, other issues that need to be taken into account. One of them concerns the contexts of the assessment situation, and the specific, more detailed construct we (as assessors or validators) want to target. And since an assessment instrument consists of tasks, the particular parameters of the tasks must also be taken into account. The entire assessment situation can be described in terms of the concept of usefulness (Bachman & Palmer 1996), where the major individual quality is construct validity, but where the qualities of reliability, authenticity and interactiveness as well as impact and practicality are also included.

A further feature of the task is the item format, and in this context the focus is on the multiple-choice format. The procedures of making sure that the multiple-choice (MC) items are valid and of high quality (the validation procedure) begin from the stage of item development and continue after the test items have been administered to the test-takers. In the following passages I will describe and discuss the different means that a test constructor or item writer has at hand for making sure that the items are of an acceptable quality.

2.1 The test context and the construct

No matter what particular language ability or skill is assessed, the context and the aim of the test are bound to influence the "final product" of the procedure of developing a test measuring that skill. This implies that even if language test developers or users know that the skill to be assessed is L2 listening comprehension, they need a lot more information in order to know what kind of a test they are to construct or implement. There are theoretical as well as practical approaches to the matter. A theoretical approach is to consider the particular "L2 listening comprehension construct(s)" that lie behind the listening ability. Ensuring that the right construct is being measured is the central issue in all assessment.

Therefore, as Buck (2001:1) points out, the first task for the test developer is to understand the construct, and then, secondly, to make a test that somehow measures that construct. A simple definition of 'construct' would be: the thing/concept/characteristic we are trying to measure (Buck 2001:1, AERA 1999: 5; Anastasi 1986: 4-5). According to Ebel & Frisbie (1991:108), the term construct refers to a psychological construct, a theoretical conceptualisation about an aspect of human behaviour that cannot be measured or observed directly. However, Fulcher & Davidson (2007: 7) states that, for a general term to become a construct, it must have the properties of being definable in such a way that it becomes measurable and so that it can have relationships with other constructs. Similarly, Linn and Miller (2005: 78) point out that *"(w)hen we interpret assessment results as a measure of a particular construct, we are implying that there is such a construct, that it differs from other constructs, and that the results provide a measure of the construct that is little influenced by extraneous factors*". A separation is made between an operational definition and a construct (Haladyna 2004: 4-6; Bachman 2004: 14). Where a construct is complex and abstract, the operational definitions are characterised by consensus and the traits defined can be objectively and directly measured. The simple things we can easily observe are operationally defined, but the most complex and prized things are not easily observable and require expert judgement. A conceptual construct definition is generally based on either a theory of language ability, or proficiency, or on the content of an instructional syllabus. The test specifications should include detailed information about types and numbers of test tasks, the amount of time allowed, and the scoring of the responses to the test tasks. This represents the operational definition of the construct (Bachman 2004: 14ff).

The skill of understanding a foreign language in its spoken mode cannot be described by a simple model, as has been established in chapter 1. Moreover, the skill – or the construct – is not static, but very much dependent on a particular context. The listening comprehension skill required for, for example, a student seeking entrance to a university where the language of instruction is foreign to him or her – it could simply be called 'L2 academic listening' – differs substantially from the L2 listening comprehension skill needed for an individual to be, for example, an efficient receptionist in an international hotel – that could be called 'L2 listening for service encounters'. If we were to assess people with these two different practical language use purposes in mind, the tests of listening comprehension would look very different from one another due to the differences in the particular listening constructs.

There are also tests of language proficiency of a "general nature" that is with no particular practical target language use domain in mind, aiming at measurement of the average level of an individual's language proficiency. Nevertheless, behind this "general language proficiency" lie philosophical as well as practical assumptions about what the concept of language proficiency covers.

In many cases today - and this is also true in the Finnish context - what we understand by general language proficiency is related to the variously defined "communicative and functional language ability". The *CEFR* as well as the foreign language curriculum within the Finnish secondary school system, the language education within higher education and the national language examinations[17] are based on this concept. Buck describes the models of communicative competence, adapted from Bachman & Palmer (1996), as *"the most widely accepted general description of language ability among language testers" (*Buck 2001: 102).

The first important step in describing a construct, in order to know what a specific test of the ability covered by that construct should measure, is followed by the second step which comprises considerations related to the limitations of the assessment instrument. In the great majority of cases, sampling for a test is necessary, since we cannot possibly cover the entire particular listening comprehension construct we have as our assessment target. Language use contexts with very specific and limited constructs may exist, but they are rare. Usually we have to look for a representative sample of the construct and as far as possible operationalize these parts or subskills.

Among crucial practical questions we find at least the following ones:

- What is the purpose of the test?
  - o Do we want an achievement test (summative assessment) related to a specific course syllabus or a language program in order to know if the learners have learnt what they should have learnt during that particular course?
  - o Do we want a diagnostic (or formative) test providing information on what the strengths and weaknesses of particular individual test-takers or of a particular group of learners are?
  - o Is it a proficiency test we need, in order to test the level of listening comprehension skill of test-takers with different language training backgrounds, for instance for a language certificate? (See further: Alderson et al. 1995: 12)

From these fundamental questions, others will follow:

- Who are the test-takers? What is their age, their background in terms of first language, nationality, schooling, language learning…? Is it a homogeneous or a heterogeneous group of people we are assessing?
- What should the test contents cover, in terms of the language proficiency level – for example according to the *CEFR* scale? This affects for example the domain, nature and size of the vocabulary, the syntactic constructions as well as the range of pragmatic and socio-linguistic elements that can or should be included in the assessment instrument.
- What language use functions or language use tasks are important to assess?

[17] These are the Matriculation Examination, the **National Certificates of Language Proficiency** and the National Language Examinations (for Civil servants), see: http://www.oph.fi/english/pageLast.asp?path=447,4699,4834,53296

- What other task parameters, in terms of input and rubric, should be considered?
- What is the test setting at large: is it an examination containing several parts or is it just the skill of listening comprehension that is assessed at one test?
    - How large groups are to be assessed at one test administration?
    - How much time can be spent on assessing the test-takers' listening comprehension abilities, and on scoring or judging their responses?
    - Are there any external restrictions decided upon concerning the test format?

This information must be explicitly included in the specifications for larger examination systems, but for smaller-scale tests, as in the case for many achievement tests for language courses, entrance tests for language institutes etc. much of the information is probably implicit and less stable.

2.2 Task parameters

When the construct and the practical aspects of the test situation are defined, there are several task parameters that need to be taken into consideration to ensure that a comprehensive and well-balanced test of listening is created. In the listening test framework of *TOEFL©* (Bejar et al. 2000) for example, the parameters are presented both for the rubric (including question, response format and rules for scoring) and the input. For the rubric, there are factors related to the instructions and the question format – what is the channel: aural or visual, the form: language or non-language and the time allotment: is there limited or unlimited processing time? As for the item-text interaction, there are variables such as the type of information requested: concrete or abstract, explicit or implicit; the type of match referring to the process (to remember, to cycle, to integrate or to generate) and to the text characteristics (the position of the necessary information and its linguistic features). The plausibility of the distractors is another feature of the item-text interaction and refers to the number of plausible distractors (see Gao & Rogers 2011) and the location of the distractors. The response format can take different forms as far as the channel (oral or written), the type (selected, limited production or extended production) and the time allotment (limited or unlimited response time) are concerned. Finally, there are the rules for scoring including criteria for correctness and the procedures for scoring (right/wrong, partial credit, rating scale).

The characteristics of the input relate to the situation prompt through the speaking participants, the topic of conversation, the setting, the purpose for listening and the possible presence of situation visuals. The text features proper concern the format of the text (the channel, the form, the possible presence of gestures and the length of the text). Linguistic features (vocabulary, phonology and syntax) and discourse features (propositional density, structure and complexity) as well as pragmatic features (the function of the speaker, the text type and the degree of planning) are decisive (Bejar et al. 2000: 26-27).

Another essential characteristic relates to the organization of the input and the questions: is the test-taker to read or to listen to the questions before listening to the spoken text? How many questions refer to one text passage? How many times is a test-taker allowed to listen to the text input?

These task parameters may be either explicitly presented and defined in the test specifications, or may exist as the test developers' and constructors' implicit knowledge only. Whatever the case, it is necessary to consider all these factors both at the stage of constructing a test, and at the test validation stage.

2.3 Test usefulness

Test usefulness is an essential concept in this context. Bachman & Palmer (1996: 18-35) describe it as a function of several qualities, *"all of which contribute in unique but interrelated ways to the overall usefulness of a given test"*. These qualities include construct validity (meaningfulness and appropriateness of the interpretations of scores), reliability (consistency of measurement), authenticity (degree of correspondence of target language use tasks to the test tasks), interactiveness (extent and type of involvement of the test-taker's individual characteristics in accomplishing a test task), impact (on society and educational systems, and upon individuals), and practicality (pertains to how the test will be developed, implemented and used).

Three principles for operationalizing the concept of usefulness are spelt out by Bachman & Palmer (1996:18). First, it is the overall usefulness of the test that is to be maximized, rather than the individual qualities that affect usefulness. Second, the individual test qualities cannot be evaluated independently, but must be evaluated in terms of their combined effect on the overall usefulness of the test. Third, test usefulness and the appropriate balance among the different qualities cannot be prescribed in general, but must be determined for each specific testing situation. These principles all need to be considered when the assessment instrument is developed, constructed and implemented. The predominance and importance of the concept of construct validity is pointed out by Fulcher & Davidson (2007: 15):

> The notion of usefulness provides an alternative way of looking at validity, but it has not been extensively used in the language testing literature. This may be because downgrading construct validity to a component of 'usefulness' has not challenged mainstream thinking since Messick.

Clearly, within the context of this research study, the concept of construct validity is the key to the quality of a measurement instrument. Nevertheless, the other components of usefulness provide important aspects of any measurement instrument in use, and they need to be – and may even be automatically so -

considered in any language assessment context. I will in the following treat the different principles behind each of the qualities.

2.3.1 Construct validity

Today testers do no longer speak so much about different types of validity as about different lines of validity evidence. Construct validity is by many scholars considered to be an umbrella term (Alderson et al. 1995) or a unifying concept (Bachman 1990: 256) covering all other traditionally separated types of validity (Cf. Anastasi 1986): content, concurrent, face, predictive and response validity; or (according to Weir (2005) and the socio-cognitive framework) context and theory-based validity and scoring validity, consequential validity and criterion-related validity. The various types of validity traditionally described are, according to Alderson et al. (1995: 171), actually different 'methods' for assessing validity. Messick (1994: 22) mentions as one reason to the view of test validity as a unified concept the many-faceted and intertwined nature of the validity issues and the difficulty of disentangling them. As is pointed out by Fulcher and Davidson (2007), validity theory is changing and evolving. They introduce the concept of pragmatic validity – based on a view that there is no such thing as an 'absolute' answer to the validity question. They say that the role of the language tester is to collect evidence to support test use and interpretation that the stakeholders (student, testers, teachers and society) accept but that this truth may change as new evidence comes to light (Fulcher & Davidson 2007: 18). If we maintain, however, that construct validity covers most of the other concepts within the large and evolving validity framework, it can be said to - as Bachman & Palmer (1996:21) put it - pertain to the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores.

The construct validity – interpreted to mean the extent to which a particular test used in a particular context measures the particular (in this case listening comprehension) construct that test constructors or examination managers have determined and described - is the most important issue in all language testing (Alderson et al. 1995: 170; AERA 1999:9; Buck 2001:1). The test users, the administrators or indeed the test test-takers want to be sure that a test of L2 listening comprehension really is a test of listening comprehension, and nothing else. Messick (1989; 1994) has provided useful terms for the two major threats to construct validity: *construct under-representation* and *construct-irrelevant variance*. Construct under-representation occurs when the operationalization of the theoretical construct is incomplete and some parts are not represented in the test, and construct-irrelevant variance occurs when the test assesses abilities that are not included in the theoretical description of the construct (Buck 2001: 94-95). Messick (1994: 21) explains the practical implications of these threats: low scores should not occur because the test is missing something relevant to the focal construct which, if present, would have permitted the affected students to reveal their competence. Furthermore, low scores should not occur because the test contains something irrelevant that interferes with the affected students' demonstration of competence.

In the context of a test of listening comprehension, the difficulty of isolating the invisible and internal skill of listening comprehension that has to be measured indirectly is obvious. How can a test-taker prove that he or she has understood pieces of spoken foreign language by other means than by demonstrating it by speaking, reading or writing? In the case of MC or other selection tasks - where the propositions are given in writing - in some cases in the first language of the test-takers, but in most cases in the target language - some L2 reading ability is required. On the other hand, if the propositions are given orally, the cognitive load on memory is very heavy, as the test-takers need to memorize – as a more or less verbatim representation – both what has been said in the spoken text passage and in the question. Where open-ended short-answer questions are used, the test-taker needs to prove his or her comprehension by writing a response, in many cases in the target language, so his or her L2 writing abilities are crucial. The use of the visual mode or actual physical actions or reactions can be a solution in some cases, but the extent to which abstract, complex information can be simplified into visuals or actions is limited. They appear to be useful mainly on lower competence levels, where there is more concrete information and where simpler relations and actions are at stake. On the other hand, depending on the particular testing context, it may not be important to isolate the skill of listening comprehension completely, as long as the other skills are not principally (or even equally) responsible for test scores, when the ability measured is listening comprehension. Moreover, the other skills or factors involved should not obscure or interfere with the listening comprehension process in any way. Construct-irrelevant variance can take two forms, "construct-irrelevant easiness" and "construct-irrelevant difficulty" (Messick 1989:34). "Construct-irrelevant easiness" occurs when extraneous clues in item or task formats permit some individuals to respond correctly or appropriately on the basis of skills that are irrelevant to the construct being assessed; "construct-irrelevant difficulty" occurs when extraneous aspects of the task make the task irrelevantly difficult for some individuals or groups. While the first type of construct-irrelevant variance causes test-takers to score higher than one would under normal circumstances, the latter causes a notably lower score.

2.3.2 Reliability

Test reliability is said to be the next most important characteristic of assessment results (after validity, see Linn & Miller 2005: 104). Reliability defines the extent to which test scores are consistent from one measurement to another – that is across different conditions in the measurement procedure (Bachman 2004: 153). Test scores can be reliable for example over different periods of time, over different samples of tasks or over different raters.

Assessment results merely provide a limited measure of performance obtained at a particular time, and if the measurement is not reasonably consistent, we cannot have complete confidence in the results. Highly reliable scores are accurate, reproducible and generalizable to other testing occasions and to other similar test instruments (Ebel & Frisbie 1991: 76).

However, there are numerous factors that influence the test results and this implies that the assessment results cannot, in fact, be perfectly consistent (Linn & Miller 2005: 105). Bachman (1990: 165; 2004: 155) presents factors that affect language test scores: the language ability, personal characteristics, random factors and the test method[18]. The language ability measured has a systematic effect on test performance – test score variance associated with this factor is thus considered to be reliable variance. Personal characteristics are age, gender, background knowledge, cognitive abilities and affective schemata (Bachman 2004: 156). The two other sources of score variance – random factors and the test method - are generally classified as sources of measurement error. The characteristics of the testing method and administration procedure have a systematic effect on test scores, since they may affect different groups of individuals differently (ibid).

Test-related factors that influence test reliability are for example the number of items in the test, the extent to which a test is homogeneous and the characteristics of individual items including their difficulty and discrimination capacity. Among examinee-related factors we can include group heterogeneity, student test-wiseness and student motivation. Administration-related factors of time limits and cheating opportunities also affect test scores (Ebel & Frisbie 1991: 88-93).

Random sources of variance include factors such as unexpected irregularities in the administration of the test, or temporary conditions of the test-takers that may affect their performance. These factors are generally unsystematic and beyond the control of the test administrators.

There are several methods of estimating reliability, but in the typical case with only one set of measures during one test event, we must depend on the internal consistency of the one test: how consistently is the skill measured within the test by means of the test items at hand? Within classical test theory (CTT) the statistical means of split-half reliability or the Kuder-Richardson formula can be used to establish the reliability of assessment scores, particularly with traditional items that are scored "right" or "wrong" (Linn & Miller 2004:111).

Within the framework of item response theory (IRT)[19], reliability is treated essentially as a precision of measurement. Contrary to CTT, when a particular IRT model fits the data, the item parameter estimates are independent of the particular group of test-takers who have taken the items. IRT estimates of measurement precision are independent of the particular group of test-takers who have taken the item (Bachman 2004: 190).

Although it is usually not possible to achieve a perfectly reliable test, and no test is free of measurement error (Linn 2006: 36) the causes of unsystematic variation should be reduced to a minimum. Consequently, the test should be implemented and marked consistently, the instructions should be clear and no item should be ambiguous (Alderson et al. 1995: 87).

[18] See also the concept of 'trait-method unit' in chapter 2.4.

[19] IRT is described and discussed in chapter 2.6.3.

As far as the relationship between validity and reliability is concerned, it is pointed out that reliability is a necessary but not a sufficient condition for validity: reliability provides the consistency that makes validity possible (Linn & Miller 2005: 105). Weir (2005: 22-24) proposes to use the term scoring validity as a superordinate for all aspects of reliability. He points out that it seems sensible to seek to enhance a test's scoring validity as far as possible without compromising the other validities (Weir's socio-cognitive framework is described in chapter 2.6).

2.3.3 Authenticity and interactiveness

Messick (1994: 21), speaking of the authenticity and directness often associated with performance assessment, points out that authenticity requires evidence that the test is not unduly narrow because of missing construct variance. Directness in turn requires evidence that the test is not unduly broad because of added method variance. The claims of authenticity and directness are thus related to construct validity and need to be supported by empirical evidence (See also Table 6 referring to Takala 1998).

Rost (2002: 123) maintains that the issue of authenticity is one of the most controversial issues in the teaching of listening. He contrasts authenticity – which can be interpreted as any and all language that has actually been used by native speakers for any ″real purpose″ – and genuineness, which refers to features of colloquial style of ″real-time planning″ that characterize every day spoken discourse. Some of the features mentioned by Rost (2002: 124) are: natural speed; natural phonological phenomenon, pauses and intonation, use of reduction, assimilation and elision; high-frequency vocabulary and colloquialisms; hesitations, false starts, self-corrections and orientation of the speech towards a ″live″ listener.

Listener roles in discourse are closely bound to authenticity and interaction. Rost (2002) points out a well-established fact in pragmatics: the closer a participant is to the ″control centre″ of an interaction, the more immediate is the purpose for the interaction, and the more ″authentic″ and meaningful the discourse becomes. A collaborative – non-collaborative dimension is presented in Rost (2002: 124) and Rost (1990: 5). Closest to the centre of interaction is the participant, followed by the addressee. Then come the auditor, overhearer and finally the judge. In a listening test, a test-taker could be placed in the auditor or overhearer category, which means that he or she is at some distance from the centre of interaction (Cf. Flowerdew and Miller 2005: 89 on the social dimension in listening).

Interactiveness and authenticity are interrelated in that these characteristics concern the contents of the test and the test-taker's processing related to real-life language use: on the basis of test performance, how well can we assume that a test-taker will exhibit the same nature and level of the ability in real target language use (TLU) situations as he or she exhibits in the test? If we have a limited and concrete construct, it is much easier to assume that the "behavior" demonstrated in a test situation is the same in a target-language use situation. If a test-taker understands the questions of a target-language visitor at a reception desk at a hotel in a simulated test situation, he or she is likely to understand the questions in a real language use situation as well. However, the construct is rarely that limited, and it is even more unlikely that a simulated situation can be used for a test event - at least in traditional, large-scale and high-stakes test contexts. Even for a text used as the basis for a test of listening comprehension, genuineness is difficult to reach, for various reasons related to the need to control or to adjust the input. We have to accept that a test is a test and in some ways a somewhat unnatural language use situation and far from the centre of interaction. However, the test needs to capture the test-taker's listening comprehension ability and the underlying processes involved in a test-taking situation should be as closely similar as possible to the processes activated in a target language use situation, so that inferences can be made to other similar situations where the ability is at stake (see Bachman & Palmer 1996: 23-29).

2.3.4 Impact and washback

The impact of a particular test varies greatly, but the basic assumption is that the more high-stakes and the more large-scale a test can be considered to be, the greater the impact on individuals and society. However, Alderson (in Chang et al. 2004)[20] points out that the concept of washback is a complex phenomenon, far from a simple case of tests having just negative (or positive) impact on teaching. After years of research[21] we know that tests have more impact on the content of teaching and the materials than on the teacher's methodology. We know that high-stakes tests have more impact than low-stakes tests – but how are we to define the stakes, as they are very individual?

In the context of a national school-leaving examination, there is potentially great influence on the teaching in the schools where the future test test-takers learn the particular subject, in this case a foreign language, French. The nature of the test, its contents as far as the linguistic elements are concerned, as well as the test format, have effect on what the teachers teach. There is often a triangular effect as illustrated in Figure 3:

[20] In: *Washback in Language Testing*, 2004 ed. Cheng et al.: Foreword: ix)

[21] For example in 1993 Wall & Alderson noted the absence of serious empirical research – now there is a slow accumulation of research.

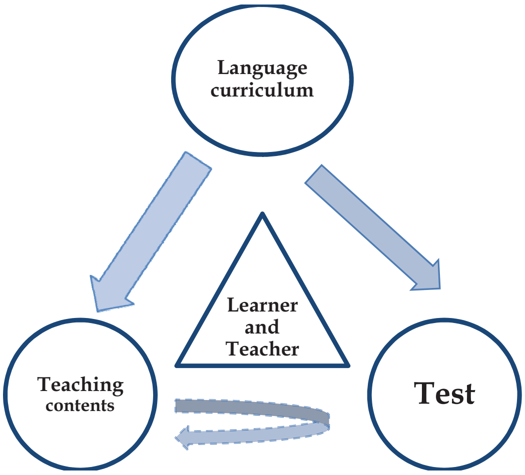


FIGURE 3 The potential impact of a test

Content-wise, the teaching as well as the test should follow the principles set out by the language curriculum: if communicative language competence lies as a basis for the curriculum, this provides the framework for what should be taught, and for what should be assessed. There is obviously space for much selection and variation within this framework, and the curricula vary in how specific they are in the operationalization of the framework: what exact lexical, syntactic and functional elements are considered essential. The more explicit the curricula, the more uniform the teaching, and the easier it is – in principle - for the test developer to construct a content-valid test, or a test that is likely to contain the same elements that are covered by the teaching. However, it is not only with respect to the contents, but also when it comes to the test format that the influence on teaching can be important. "Negative washback" may result in cases where a teacher considers it necessary to teach to the test, putting main focus on preparing the learners for the assessment event, and thus use less time for the ultimate goal, namely to also be able to use the language in various real-language use situations outside the test. "Positive washback" would occur in cases where the teaching of a language was poor – according to some generally accepted criteria - and the test would be a factor motivating the teacher to improve his or her quality of teaching. A positive washback example with a test of listening comprehension would be in contexts where the focus is on written language only – on reading, writing and grammar – but where the external test of listening comprehension encourages the teacher to include elements of oral communication. *"What is assessed becomes valued, which becomes what is taught"* (Cheng et al. 2004:3) is a statement with a lot of truth in it.

The central goal of a test remains, however: it is administered to measure an ability and by means of the test we can obtain information on the basis of which we make decisions. Importantly, the test is usually not in itself a pedagogic task aiming at learning[22]. However, with the washback effect on teaching and learning situations, caused for example by the knowledge of the coming high-stakes test after a language course, the teacher in charge of the teaching as well as the learner are bound to want to teach and to learn the material supposedly covered by the test. The test tasks are not necessarily directly applicable to real-life target language use situations, so the challenge for the teacher is to balance between including in his or her teaching tasks that are applicable to real life, as well as tasks that may be included in the test. It is usually a question of honor and will be considered a sign of the quality of the teaching by an individual teacher to be able to show that his or her students have done well in an important examination. The likelihood of negative washback can be reduced by the test constructors if they develop a test of as high a quality as possible, taking a wide range of aspects into account (cf. Weir 2005: 134).

Larger impact for the society in general may indeed occur on different levels: from the creation of manuals or the marketing of courses with the goal of preparing for the test, to governmental decisions based on some feature of a test. For an individual, the impact is obvious, ranging from whatever the stage of test preparation implies, to the feedback given on the test success to, most importantly, the decisions made on the basis of the test score on for instance entrance to university, migration or a job (See: Bachman 1996: 31).

Alderson (2004, in Cheng et al., Foreword: xi.) questions the responsibility of the test developer in the validation situation:

> In the current views of the nature of test validity, the "Messickian view" [see for example Messick 1994] of construct validity, it is commonplace to assert the need for test validation to include a consideration of the consequences of test use. […] I have serious problems with this view of a test's influence, not only because it is now clear that washback is brought about by people in the classrooms, not by test developers, but also because it is clearly the case that there is only so much that test developers can do to influence how people might prepare students for their test "

However, Cheng et al. (2004) point out that researchers have started to pay attention to the specific educational contexts and testing cultures with which different types of tests are being used for different purposes, so that implications and recommendations can be made available to education and testing organizations in many parts of the world.

2.3.5 Practicality

In most large-scale examination contexts, there are practical restrictions on several levels to be considered. It all comes down to restricted resources: "lack" of time and money. Therefore, there are restrictions as to the test planning and construction stage: how much time and how large a pool of specialized test developers can be used? Is there any possibility of piloting the test items and of analyzing the test results both from the piloting stage and the post-test stage? As large numbers of test-takers are tested, there are restrictions as to the test

[22] See, however, the description of Dynamic Assessment (ch 1.1 above)

methods or formats used, also influencing the authenticity of the test items, the length of and the conditions for implementing the test. One example is the decision from the Finnish Ministry of Education not to include an oral subtest (the skill of speaking) in the test battery for the foreign language Matriculation Examination. The debate is heated on how this will influence the contents of teaching, which should be better adapted to follow the principles of the *CEFR*. The oral skill is supposed to be included by means of focusing on that skill within one particular foreign language course module at the upper-secondary level, but also possibly within that module measured by an external oral test produced by the Board of Education, which has so far been a voluntary test for the upper-secondary students who want a separate certificate on that subskill.

The skill of listening comprehension is measured in the foreign language Matriculation Examination by a test where the input comes from a tape recording, and where one large part of the test items are selection-type items, either MC or true-false items, which can be scored by machine and do not need human input in shape of external independent assessors. In case of open-ended questions, of which there is today usually a certain amount in each test, individual teachers first need to give a first evaluation of their own students' performance which is then checked by centrally appointed judges. These open-ended short-answer questions surely contribute to making the test more construct-relevant, as different parts of the construct are probably pinned down with using different types of test items. Moreover, there is the possibility of, for some items, giving partial scores, which better reflects the way the test-taker's ability is shaped as a continuum. However, compared with the situation where MC questions are used, more resources and more subjective judgment need to be involved in the process at the stage of assessing the test-takers' responses.

2.3.6 Relative importance of different test qualities

The balancing of the relative involvement of different test qualities is an important issue that test constructors and test administrator need to be aware of and to take responsibility for. It seems obvious, however, that the characteristic of construct validity should be the most important quality and if compromises have to be made, these should not be made at the expense of the validity of the test scores. Accordingly, Weir (2005: 49) points out that practicality issues should not be considered before sufficient validity evidence is available to justify interpreting test scores as an acceptable indication of the control of an underlying construct. He stresses the risk of practicality intruding on the test in such a way that it will not be assessing what we want it to assess.

2.4 Test and item formats

In the preceding passages I have tried to show that many aspects covered by the concept of usefulness influence the format of a test, and, conversely, the test format largely influences the quality and the usefulness of a test. Therefore, in the following, I will contrast features of the MC format with other possible test formats.

The choice of an item format has many implications and presents many problems to the test developer, being one of the most fundamental steps in the design of any test. A fundamental principle in the choice of an item format is that measuring the content and the cognitive process should be the chief concern (Haladyna 2004: 41-42). The item format (or the test rubric, using the concept of Bejar et al. 2000) can be defined as a device for obtaining a student response. Its components are a) a question or command to the test-taker, b) some conditions governing the response and c) a scoring procedure. The large effect that the item format has on the cognitive processes of test-takers was evidenced in a study conducted by Rupp et al. (2006). They analysed the responses given by ten test-takers during interviews related to MC reading items[23], and found out that:

- There exists multiple different representations of the construct of "reading comprehension" that are revealed through the characteristics of the items.
- Learners view responding to MC questions as a problem-solving task rather than a comprehension task.
- Learners select a variety of unconditional and conditional response strategies to deliberately select choices.
- Learners combine a variety of mental resources interactively when determining an appropriate choice (Rupp et al. 2006: 441).

Good reading tests should employ a number of different formats, making sense since in real-life reading, readers typically respond to texts in a variety of different ways (Alderson 2000: 206; Cf. Haladyna 2004: 42). This aspect of authenticity will apply to listening tests as well. An added reason to include several task types within a test is that possible method effects and biases are weakened by the use of several test formats instead of only one. Messick (1994: 22), in an article focusing on performance assessment in general, recommends that assessment batteries represent a mix of efficient structured exercises and less structured open-ended tasks. Linn & Miller (2005: 165) maintains that each type of test item has its own unique characteristics, uses, advantages, limitations and rules for construction. An important concept here is the trait-method unit[24],

[23] Reading comprehension bears some similarities with listening comprehension. Apart from both being "receptive skills" as contrasted with the productive skills of writing and speaking, these skills are often measured with similar test formats, especially at the present target ability level. Therefore, some studies and findings on the assessment of reading comprehension are relevant to the present study, and thus referred to here.

[24] Campbell and Fiske (1959) in the Psychological Bulletin stated that each *"test or task employed for measurement purposes is a trait-method unit, a union of a particular trait content with measurement procedures not specific to that content."* Moreover, they established that *"the systematic variance among test scores can be due to responses to the measurement features as well as responses to the trait content"* (reprint of the article in Ward,

which implies that a test-taker's performance is a function of two variables: the test-taker's language ability and the test method (See for example Bachman 1990: 225; Yi'an 1998: 21). The assumption is that a test format may influence the processes a test-taker makes use of in a test situation, and the question is whether these processes are compatible with the particular listening construct that the test constructors and users have targeted (Cf. the notion of interactiveness). This is related to both the reliability of test scores obtained on particular test items and the validity of the uses of these scores.

2.4.1 Multiple-choice items and other formats

There is frequent and sometimes harsh criticism raised concerning the construct validity of MC tests in general, the most serious being that examinees can pick the correct answer without comprehending the text (Freedle & Kostin 1999:3). Nevertheless, the MC item format persists as a frequently used and sometimes as the only type of item included in standardized language tests. Haladyna (2004: ix) maintains that despite attack on MC testing, it has thrived in recent years. Apart from the opinion that it appears to be a format that can be used for testing most kinds of knowledge and comprehension on various levels (Cf. Ebel & Frisbie 1991: 154, Haladyna 2004: 6; Linn & Miller 2005: 194), it is objectively and easily scored, even by machine in case of large-scale assessments. Buck (2001: 146) establishes that although complex and difficult to make, MC items can be used to test a variety of listening sub-skills: from understanding at the most explicit literal level, through combining information from different parts of the text, making pragmatic inferences, understanding implicit meanings, to summarizing and synthesizing extensive sections of text. Even though the most important constructs are not best measured with MC item formats, MC tests still play a role in measuring important aspects of many constructs (Haladyna 2004: 6).

Bailey (1998: 130) whose approach (based largely on Oller, 1979) is to caution teachers of the dilemmas presented by the MC format, lists the possible reasons for applying the format: it is fast, easy and economical to score; it can be scored objectively (giving the appearance of being fairer and more reliable); it looks like a test, and, in comparison with true-false items, it reduces the chances of learners guessing the correct answer. However, Bailey (1998) establishes that a great deal of subjective judgment goes into the development of MC items and that writing good MC items is extremely laborious. Her main warning concerns the negative washback of using the MC format (ibid: 131). In their discussion of the advantages and disadvantages of the format, Mendelsohn & Rubin (1995: 43) make the point that as MC items require a minimal amount of time to complete, a test can include many items, enhancing its reliability. The fact that MC items minimize the confounding of listening with speaking or writing, speaks for

A.W., H.W Stoker and M Murrey-Ward (eds.). Educational Measurement: origins, theories and explications, vol.1 (pp. 225-234).

their use. MC items are practical in situations that require testing of large numbers of individuals. Disadvantages that speak against their use are the invitation to guessing, the possible problem of creating plausible distractors (see Gao & Rogers 2011: 98) for some important parts of a passage and the fact that good MC items are difficult to write.

In a comparison of the MC test type with other objective tests Linn & Miller (2005: 195-196) reach the conclusion that it is easier to construct high-quality test items in MC format than in any of the other forms even if this does not mean that good MC items can be constructed without effort. But for a given amount of effort, they claim that MC items will tend to be of a higher quality than short-answer[25], true-false[26], or matching-type items[27] in the same area. Comparing the MC format with the short-answer question type, the researchers maintain that the ambiguity and vagueness that are frequently present in the short-answer item can be avoided because the alternatives better structure the situation.

However, if the short-answer questions are carefully formulated, as Weir (2005) points out, a test-taker's response can be brief and thus a large number of questions may be set. It is important to ensure that the questions are phrased in simple language, as it may not be possible to have questions in the test-takers' L1. Trialing should be used to make sure that the short-answer questions are unambiguous, and sufficiently focused, and to determine the range of alternative correct answers (Weir 2005: 125). One advantage of the short-answer format is the possibility of giving partial scores in cases where for example the test-taker has given one part of a two-part necessary information (NI) as his or her response, or where it is considered that a partially correct response does not merit a total loss of a score. This would better reflect the nature of the listening ability, representing a continuum more than an on-off situation.

Two disadvantages of the open-ended test-format for the test of listening comprehension are, on one hand, that it demands writing skills of the test-taker and, on the other hand, that it may be difficult to objectively determine what constitutes an acceptable interpretation and an adequate response from the test-taker (Weir 2005: 140). If there seems to be next to an infinite number of possible responses, there is something wrong with the question. Information transfer – where the information transmitted verbally is transferred to a non-verbal form, for example by labeling a diagram, completing a chart etc. - is mentioned as a useful variant of the short-answer questions, as here the verbal part of the answers can be kept to a minimum. However, there is the risk of the non-verbal task further complicating the processes – in a test of listening a test-taker may understand the text but not what the transfer task requires.

[25] Short-answer questions are generically those that require the test-takers to write down answers in spaces provided on the question paper (Weir 2005:124).

[26] A true-false item is a type of selected response where the test-taker has to indicate whether a series of statements are true or false in relation to a text (UCLES, Multilingual glossary of language testing terms, Studies in language testing 6).

[27] A matching-type item involves bringing together elements from two separate lists (ibid.).

A weakness of the true-false item is that students can receive credit just for knowing that a statement is incorrect, whereas in a multiple-choice item, they also have to know what is correct. According to Haladyna (2004: 77) the true-false (TF) format has been well established for classroom assessment but seldom used in standardized testing programs. There are problems related to the use of the TF format, such as its large error component due to guessing, and experts rarely recommend its use. Where there are recommendations for writing TF items, they concern the balancing of true and false statements, the use of simple declarative sentences, the use of internal comparison clearly stated in the item (unambiguous items) and the use of MC items as a basis for writing TF-items.

Rantanen (2003: 60) mentions the multiple true-false format (MTF), where there is a question followed by several options that can be independently true or false. According to studies the test reliability could be enhanced in comparison to MC items. The results of the study by Kreiter & Frisbie (1989) also provide support for the use of MTF tests as an alternative to MC tests in the measurement of achievement: the MTF measures were shown to yield higher reliabilities, lower adjusted means, and higher response rates when compared with those of MC (cf. Haladyna 2004).

As far as the issues of interactiveness - that is authenticity as a function of the processes activated for the different test formats - is concerned, one approach is to imagine different real-life TLU situations that involve listening comprehension. It is partly related to the proficiency level of the learner and varies according to the current communicative situation. We can assume that the situation does not involve mere "listening for pleasure" but a communicative task of some kind.

A typical "high-level" language situation and task would be listening to a lecture and taking notes. These notes would then be used to create a new "knowledge bank" – explicitly or implicitly. The process would first involve understanding the more or less explicitly stated meaning of the spoken language of the lecture (with all its different characteristics), then forming a mental representation and deciding on what is relevant and important new information, worth noting down and taking in. From the point of view of interactiveness we could imagine a test-task where a test-taker would fill in missing key information in a text, or answer short-answer questions on the contents of the lecture. The drawback would be the interference/influence of the test-taker's background knowledge on the lecture subject and his or her writing skills - a possible case of construct-irrelevant variance. And in addition, there is the subjectivity of the assessor's judgment.

Another real-life task would be to take in information and make a decision on the basis of that information: for example a situation where the listener-learner takes L2 advice on what gifts to buy for Christmas for different people. This may involve a fairly important amount of interactiveness, but could be limited to a situation where the listener would not interact very much, but write down possible gifts for a list of people. Besides the listening process this would involve a proper socio-linguistic construct – a code of practice associated with Christmas traditions.

The advantage for the test is the fact that the level of language could easily be adapted to different learner levels. The task could consist of matching the spoken text to a list of pictures, or concepts – a natural task imitating for example the task of selecting a gift among what is offered in a web catalogue. Here the interference from reading or writing skills is minimal.

A third example from a TLU context could involve more oral interaction – a situation where the listener also takes the role of a speaker in a dialogue. This could for instance take place in a virtual hotel, where the listener listens to a question asked by a "hotel employee" and responds in speaking. Here the information is probably personal, so no important subject knowledge is needed. This TLU task would transfer rather easily to a test situation, provided that the orally answered information does not put too much of a cognitive load to the task for the test-taker, causing construct-irrelevant variance.

These task examples can be related to the comparison of different aspects pertaining to authentic, performance or direct assessment (Takala 1998; Table 6 in the present thesis) as well as to the model of information-processing at a MC test by Jamieson et al. (2000) presented in Table 7 below. The nature of the processing stages differs according to the task type and the item format, and so do the variables affecting the process. This is an important validity issue that has to be considered: what characteristics and variables in the process do we want to include in the construct that is targeted at one specific testing context?

2.4.2 The importance of the quality of the MC item

As is unanimously concluded, doubtless, the task of constructing good MC items is a complicated one. That may be a disadvantage speaking against their use, unless the ease of scoring makes up for the time and effort put into the construction process. Alderson (2000: 212) says that the construction of MC questions is a very skilled and time-consuming business and the task of writing plausible but incorrect options that will attract the weaker reader but not the better reader is far from easy. Therefore - due to their complexity and their unpredictability - MC items ought to be pre-tested before being used in any high-stakes assessment (Buck 2001: 142).

The worries as to the quality of MC items are related to the validity of the test – to the question whether items aimed at measuring listening comprehension actually measure that construct. The issue at stake relates to the processes activated at a test consisting of MC tasks. As listening comprehension is used to make sense of the linguistic input in the light of the purpose for listening, the questions and optional answers in a MC listening comprehension test serve as the imposed and therefore shared purpose for listening and exert an impact on the listeners' listening process (Yi'an 1998: 36). Consequently, even if in the ideal case, *"a language-comprehension test 'should' assess primarily the difficulty of the text itself – the item structure itself should only be an incidental device for assessing test difficulty"* (Freedle & Kostin 1999: 3) – the importance of the nature of the task, of the question and of the options is indisputable. As far as the factors of difficulty in tests of reading comprehension are concerned, it appears to be difficult to distinguish between item effects (covering question or task characteristics) and text passage effects, as the two interact (Alderson 2004: 86). The estimation of what contributes to item difficulty is an important area for testing research. The situation is certainly the same for listening tests as well.

The MC task itself requires certain processes that may not be common to other comprehension situations, related to task performance skills and problem solving skills (Buck et al. 1996: 612). Thus, many scholars believe that the MC items have a strong method effect (Brindley 1998) and that they make considerable processing demands on the test-taker (Hansen & Jensen 1994). They can force test-takers to re-adjust their interpretation if it does not agree with the options (Nissan et al. 1996 in Buck 2001: 143). Moreover, the MC test method encourages students to consider alternatives they would not otherwise have considered – thus the technique tricks the unwary into making incorrect interpretations they might not otherwise have made (Alderson & al. 1995: 45). In cases where the test-takers are from different non-L2 backgrounds, and the question and options are given in the L2, the test is also measuring a certain amount of L2 reading comprehension. In Alderson's experience, some comprehension items do not test what they are intended to test: items may turn out to be testing background knowledge. It is unfortunately easy to write items that can be answered without any reference to the reading or listening passage (Alderson 1995: 50).

One of the more serious difficulties associated with MC questions is that the tester does not know why the test-taker responded the way she did (Alderson 2000). He or she may have simply guessed at his or her choice, or he or she may have a totally different reason in mind from that which the test developer intended when writing the item. The test-taker may simply have employed test-taking strategies to eliminate implausible choices, and has been left with only one choice. As Alderson (2000) says, researchers can explore the processes test-takers engage in when validating their tests, but he maintains that there is no guarantee that any given test-taker will in fact use processes that have been shown to be commonly used. Alderson (2000: 212) speculates that in the situation where test-takers were required to give their reasons for making their choice as well, the problem might be lessened, but points out that the practical advantage of MC questions in terms of marking would in that case be lost.

An interesting and illustrating analysis of the model of information processing (goal-process system) at a multiple-choice test of listening (within the TOEFL test battery) conducted by Jamieson et al. (2000) is presented in Rost (2002: 177). Table 7 shows the various activities and challenges associated with the multifaceted process:

TABLE 7 A model of information processing (goal-process system) at a multiple-choice test of listening (Jamieson et al. 2000)

Stages	Goal	Process	Variables that affect the process
1: Listening to the stimulus	Listen to the stimulus and remember information in order to answer each question following the stimulus	Represent in working memory information in the stimulus regarded as important	Stimulus variables: length of lecture, syntactic complexity, density of information, lexical difficulty Listener variables: knowledge of the context of the task, knowledge of the language, attention, working memory capacity, background knowledge.
2. Listening to/ reading each question	Understand the questions	Identify the given and requested information in the question and represent in memory the requested information	Item variables: lexical difficulty, syntactic complexity, length Listener variables: (as above)
3. Searching for the correct answer	Retrieve information from stimulus that answers the question.	Search working memory for information in the stimulus that matches the information requested in the questions.	Stimulus variables: (as stage 1) Item variables: type of information, type of match, explicitness, main/supporting idea, redundancy Listener variables: (as above)
4.Identifying the correct answer	Select the correct answer from the options given	Identify an answer to the question by finding a match with the appropriate information from working memory and verifying that none of the other options is a better match.	Stimulus variables: Item variables: Listener variables: (as above)

This model gives a good clue to what the effect of the item format might be and how the potential processes differ from a situation where a test-taker is requested to give an open response to a stimulus. It must be pointed out, though, that the stages do not necessarily occur in the above order, but there is probably simultaneity or movement back and forth between them. Nevertheless, at every stage, the item including the question and the options affect the procedure. If the test-taker gets to read the question before listening, the question referring to a particular passage determines what is important in the input (the spoken text passage), and what should be represented in working memory. This occurs at the second stage at the latest, where item variables, that is, features in the possibly written questions and options, are crucial. Then through the third and fourth stages, the decoding, the understanding and the interpretation of the item are as important as the understanding of the spoken text. This illustrates well how great the effect of each test item is, and further underlines the importance of creating transparent questions that provide a clear purpose for listening.

2.4.3 Multiple-choice test-taking strategies: guessing and elimination

The two typical, even inherent, strategies employed for the MC item format are the strategies of guessing and elimination. The use of these strategies is said to be a possible threat to the validity and the reliability of the test scores. Clearly, the tester must be well aware of these factors when MC items are used. The important question is to what degree they might affect the quality of a test and whether they distort the measurement target: the skill of listening comprehension.

As Haladyna (2004: 217) points out, with the use of MC items, an element of guessing exists. It is, however, important to note that when facing MC questions of any kind, even if you do not know the response for sure, you will only rarely be in a situation where the guess you make is completely random. More often than not, you will have some background information that might, for instance, tell you which options are definitely NOT correct, which is also valuable information. Bachman & Palmer (1996: 205) relate the test-taking situation to a real language use situation, saying that if we get lost when faced with spoken discourse, we virtually never make a totally random guess at meaning; rather, we use the means at our disposal - language knowledge, topical knowledge and metacognitive strategies – to arrive at a possible understanding of the meaning. Linn & Miller (2005: 343-344) point out that problem-solving always involves a certain type of informed guessing. They compare the guesses on doubtful items with the informed guesses we make when we predict weather or judge the possible consequences of a decision.

The reasons for guessing are multidimensional. Among the factors affecting the test-taking process are the spoken text as one dimension, the task including the written question (the stem) and the three options as another and as the third dimension the test-takers with their various personalities, skills and knowledge (language and other) and experiences. All these dimensions influence the test-takers' tendency to guess at an item (cf. Bachman & Palmer 1996: 204-205, Rantanen 2003: 102, 187).

The interesting points to investigate are, first of all, **how** the strategy of guessing is used at a particular test by particular test-takers and, second, **why** they use it. These two questions are interrelated and partly dependent on each other. The situations where a test-taker relies on guessing do not just reveal something about the nature of the test format, but also - and possibly to a greater extent - about the quality of a particular test item. When considering how the MC item should be constructed, the number of options is one important factor. The larger the number of options, the less likely is a correct random guess. However, the form and contents of the options will no doubt affect the truth of this principle.

The availability of the process of elimination is sometimes regarded as a weakness of the MC format (Ebel and Frisbie 1991:156-157). However, the use of this rather demanding strategy can be considered justified, as the knowledge and ability required to properly eliminate incorrect alternatives are related to the knowledge and ability required to select the correct alternative. Few MC items are likely to be answered correctly merely by eliminating incorrect choices and often the process will actually involve comparative judgments of one alternative against another.

Solving any MC item contains the elements of guessing and elimination, depending on the level of knowledge the test-taker can rely on at a particular item (Haladyna 2004: 217). Test-takers normally have some comprehension of the point being tested. Thereby, implausible distractors can be eliminated on the basis of partial knowledge (cf. Haladyna 2004: 217; Bachman & Palmer 1996: 204; Buck 2001: 147). Therefore, even if the test-taker does not know the correct answer, he or she is inclined to favor one response over another, perhaps without even knowing exactly why.

In Cohen's (1998a) classification, educated guesses – using background knowledge or extra-textual knowledge - are seen as belonging to the category of language-use based strategies, whereas elimination is classified as a test-wiseness strategy. The process of elimination is described by Cohen (1998a: 230) as [to] select a choice not because you are sure that it is the correct answer, but because the other choices don't seem reasonable, because they seem similar or overlapping, or because the meaning is not clear to you. In the present study I will interpret elimination as a larger category still, including Ebel and Frisbie' s (1991) interpretation of elimination as making comparative judgments of the options.

2.5 Central issues on the usefulness of the test and the test format

Some of the central issues for this study are related to the broad concepts of usefulness and construct validity, and to the MC test format. First of all, we have established that the "product" consisting of listening comprehension items can only come about by means of a complicated process, where many aspects based on the principle of usefulness interrelate and influence the outcome. There are theoretical and practical restrictions as well as ethical factors to consider all through the process of developing, constructing and implementing a test of listening comprehension. As is pointed out by Bachman & Palmer (1996:18) the factors that together make up the overall usefulness must be evaluated in terms of their combined effect on the test. However, the (construct) validity is taken to be the dominating/superior factor as far as the test quality is concerned (See for instance Alderson et al. 1995: 170; AERA 1999: 9; Buck 2001: 1).

Second, an important concept is the trait-method effect that is taken here to not only explain the functioning of a test method in general, but also the processing of test-takers at individual items with varying characteristics. These characteristics are related to the risk of construct-irrelevant variance.

Third, the use of the MC format comes with problems and drawbacks related to test usefulness. The format does seem to lend itself to testing various sub-areas within the skill of L2 listening comprehension. However, the task of developing good MC items testing comprehension in a valid and reliable way is not easy, and consequently, resources should be put into the process of validating the test and the separate items it consists of. The statistical data obtained after the administration of a test or preferably a pre-test gives information about where potential flaws in the items may be. However, in order to get deeper into the reasons behind the functioning or the non-functioning of separate items or testlets, the statistically acquired information should be completed with more qualitative information. In the following chapter (2.6) I shall present different means that can be used as parts of the validation procedure, that is, tools that can be used to increase the likelihood that the test is a valid and reliable measure of the particular skill, in this case the skill of foreign language listening comprehension.

2.6 Multiple-choice item development and validation

How can we claim and prove that a test does measure what it purports to measure? The process of construct validation, of providing evidence for 'the adequacy of a test as a measure of the characteristic it is interpreted to assess', is a complex and continuous undertaking, an ongoing process of accumulating theoretical and empirical evidence (Bachman 1990: 270-1, Buck 2001: 195). It should be pointed out, however, that what is possible to carry out within the framework of a research project, may not be feasible in the context of most test construction situations. What would be an ideal case of validation procedures for obtaining well-functioning test items may not be possible to undertake in practice. However, one can assume that it will be in the interest of every examination board and every test development authority to do their best to produce assessment instruments of high quality, as well as to prove their quality, by means of using as thorough validation procedures as possible.

Validity is multifaceted, and complementary aspects of evidence are needed to support claims for the validity of the scores on a test. The more validity evidence that can be gathered for a test the better. Many lines of evidence can contribute to an understanding of the construct meaning of test scores (Cf. Weir 2005: 13; Alderson 1995; AERA 1999: 5; Anastasi 1986: 3). As Weir (2005) points out, the different types of evidence are not alternatives but complementary aspects of an evidential basis for test interpretation. Kane (1992: 534) speaks of the advantages of an argument-based approach to validation, where no specific kind of validity evidence is preferable to any other kind of evidence, but which requires that the interpretative argument is stated as clearly as possible and that the validity evidence should address the plausibility of the specific interpretative argument being proposed. Messick (1994: 22) says that the interpretation and use of performance assessments, like all assessments, should be validated in terms of content, substantive, structural, external, generalizability, and consequential aspects of construct validity.

The result of this process of construct validation will be 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 (Bachman 1990: 270-1). Validation can thus be seen as a form of evaluation where a variety of quantitative and qualitative methodologies are used to generate evidence to support inferences from test scores - the empirical testing of hypothesized relationships between test scores and abilities (AERA 1999: 9; Bachman 1990: 258, Weir 2005: 15).

The task for the test evaluator or a researcher in a validation study is to collect as much proof and counterproof for and against the intended interpretation of test scores as possible. Bachman (2004: 264) talks about claims and counterclaims about factors that are likely to affect performance on the test:

> The central claim […] is that performance on the test task is affected primarily by the area(s) of language ability we want to measure. We also claim, even if implicitly, that test performance is not affected in any important way by factors other than the ability we want to measure. The potential effects of other factors that may affect performance on this test need to be articulated as counterclaims in our validation argument.

Plausible rival hypotheses - counterclaims - can often be generated by considering whether a test measures less or more than its proposed construct. Such concerns are referred to as construct underrepresentation and construct-irrelevant variance (Messick 1989; AERA 1999: 10). Validation involves careful attention to possible distortions in meaning arising from inadequate representation of the construct as well as to aspects of measurement such as test format, administration conditions, or language level that may materially limit or qualify the interpretation of test scores (AERA 1999: 10).

The following sources of validity evidence are presented in the *Standards for educational and psychological measurement* (AERA 1999: 11-16), widely referred to in the field of language assessment:

1) Analysis of the relationship between a test's content and the construct it is intended to measure.
2) Theoretical and empirical analyses of the processes of test-takers.
3) Analyses of the internal structure of a test
4) Analyses of the relationship of test scores to variables external to the test, including measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs, and tests measuring related or different constructs.
5) Analysis of the intended and unintended consequences of test use[28].

[28] An issue that has been receiving more attention in recent years (AERA 1999); cf. Chang et al. (2004) on the issue of washback.

For the purpose of establishing theory-based validity evidence within the "socio-cognitive framework", Weir (2005: 233) has drawn a table outlining a variety of procedures that may be used to establish what is happening when test-takers are actually performing on the test tasks. He divides the procedures into three stages (cf. Table 8 below).

TABLE 8 Procedures for establishing quality-based evidence (Weir 2005: 233)

Stage 1A	Qualitative expert judgement of items
Stage 1B	Qualitative introspection/retrospection by test takers (think aloud/interview/ questionnaire) to validate strategies and skills and the conditions under which the test tasks are performed.
Stage 2	Quantitative and qualitative analysis of test performances Basic descriptive statistics Correlation, Factor analyses of test results, *t*-tests, Multi-faceted Rasch Qualitative discourse analysis of test performances in productive tasks

The three first sources of validity evidence provided by AERA (1999) are treated in our present research context. These parallel the procedures categorized by Weir (2005) to cover what he calls theory-based validity evidence. According to *the Standards for educational and psychological testing* (AERA 1999: 11) important validity evidence can be obtained from an analysis of the relationship between a test's content and the construct it is intended to measure. Test content refers to the themes, wording and format of the items, tasks or questions on a test, as well as the guidelines for procedures regarding administration and scoring. Evidence based on test content can include logical or empirical analyses of the adequacy with which the test content represents the content domain and can come from expert judgments of the relationship between parts of the test and the construct.

Both quantitative and qualitative features of the test contents should be considered. The hypothesized impact on the difficulty of an item is looked at through different features in the approach put forth by Rupp et al. (2001). Quantitative features are found in Table 9 in variables 1, 2, 3, 4, 5, 11 and 12. Mainly qualitative features of the test contents are taken into account by analysing features 6, 7, 8, 9 and 10. This seems to provide a very good coverage of characteristics having an effect on the test contents with respect to item difficulty.

TABLE 9 Independent text, item and text & item variables and their hypothesized impact on item difficulty (From: Rupp, Garcia & Jamieson 2001)

Variable label	Description	Hypothesized Impact: Items are more difficult if…	Relation to construct: because…
1) Word count	Total nr of words in the text	The word count is higher	There is more text to process
2) Sentence length	Average nr of words per sentence in the text	The average sentence is longer	Longer sentences related to syntactic complexity, harder to understand
3) Type-token ratio	Index of the lexical richness of the text	The type-token ratio is higher	There is a higher information load, and the text is more difficult to process
4)Information density	Index of information density of the text	The information density is higher	The information is highly condensed and harder to process
5) Lexical overlap (distractors-correct answer)	Nr of distractors that have at least one content word in common with the correct answer	The lexical overlap between the distractors and the key increases	The differences among the options require finer discrimination
6) Item type	The correct answer requires a detail, main idea or gist, prediction, understanding relations	The correct answer requires understanding relations	The cognitive processes required for each item type are progressively more complex
7) Type of information	Complexity of requested information from concrete to abstract	The type of requested information is more abstract	Abstract information is more difficult to recall than concrete
8) Type of match	Index of processes used to "match" question to text and to select option	The required processing strategies are more complex, and there are more features to search on	The number and complexity of cognitive operations increases
9)Directness of information	Explicit versus implicit	The requested information is only implicitly provided	Inferencing is more cognitively demanding than recognizing information
10) Location of information	The section of the text (first/second/ last) with the last mention of the requested info	The information is located earlier in the text	The information may no longer be in one's short-term memory
11) Nr of plausible distractors	Nr of distractors that are plausible given the situation described in the text	The nr of plausible distractors increases	Finer discrimination will be needed to identify the requested information
12) Lexical overlap (text-correct answer)	Index measuring the amount of identical content words between correct answer and text	The lexical overlap between the text and the correct answer is lower	Lack of key

According to Weir (2005: 19), the term context validity would better account for the social dimension of language use:

> Context validity is concerned with the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample. This coverage relates to linguistic and interlocutor demands made by the task(s) as well as the conditions under which the task is performed arising from both the task itself and its administrative setting.

Related to the content analysis, analyses of the internal structure of a test can indicate the degree to which the relationship among test items and test components conform to the construct on which the proposed test score interpretations are based. As Buck (2001: 116) points out, it is important to take account of the fact that each individual task may only operationalize part of the construct, but taken together the tasks need to represent the whole construct. Analyses of the more practical issues related to test conditions, such as its setting and administration and their affect on the test-taker performance are also essential parts of context validity.

Some methods for studying the internal structure of tests have been devised to show whether particular items may function differently for identifiable subgroups of examinees. Differential item functioning occurs when different groups of examinees with similar overall ability, or similar status on an appropriate criterion, have, on average, systematically different responses to a particular item. However, differential item functioning need not always be a flaw or a weakness. Subsets of items that have a specific characteristic in common (e.g. specific content, task representation) may function differently for different groups of similarly scoring examinees. This indicates some kind of multidimensionality that may either be unexpected or it may conform to the test framework (AERA 1999: 13).

Evidence based on response processes comes from analyses of individual responses and can give valuable information on test takers' performance strategies and processes. There is strong proof from earlier studies that test takers' performance varies not only because of differences in the ability measured, but also because of differences in the processes or strategies they use when responding to test tasks. Different test takers may thus get the same task right although they use different processes, or different test takers may get the same task wrong for different reasons (Bachman 2004: 276). Therefore, studies with examinees from different subgroups can provide evidence of differences in meaning of test scores across different subgroups, as well as of the extent to which capabilities irrelevant to the construct may be differentially influencing their performance (AERA: 12). The concern that needs to be addressed with respect to validation, then, is the extent to which different processes are engaged on the same task by different test-takers (Bachman 2004:276).

As we are not able to observe directly the cognitive processes that test takers use when responding to test tasks, we need to rely on indirect evidence. One way to collect empirical evidence about processes used in taking tests is to ask test takers to provide a verbal report on the processes they use. While verbal report analysis typically involves the qualitative analysis and description of the processes that test takers report, it is also possible to group the processes reported into categories of specific strategies. Bachman (2004: 278) points out that the use of verbal protocol analysis has provided many valuable insights into the ways in which test takers process different kinds of language test tasks. He thus considers the methodology an indispensable tool for collecting information about test performance as part of the test try-out phase before tests are used operationally to make decisions.

An important point of view is provided by Haladyna (2004: 262). He says that as the test item is the most basic unit of measurement, it matters greatly that we address the issue of validity evidence at the item and item response levels. Therefore, even if the pool of items as a whole matters in judging the quality of the test as an assessment instrument of the skill of listening comprehension, it is important to take the processing at each separate item into account.

In validating a specific test score interpretation or use, one body of evidence comes from item development (Haladyna 2004: 1; Anastasi 1986: 3). Another body of evidence resides with studies of item responses. Item development links to validity in the following manner:

> We can define what an item is supposed to measure and the type of cognitive behaviour it elicits. We can write the item, which is the explication step in construct validation, and we can study the responses to the item to determine whether it behaves the way we think it should behave (Haladyna 2004: 18).

These three steps of construct validation are presented in Table 10.

TABLE 10 Three steps of construct validation (From Haladyna 2004: 18)

Three steps	*Test Score*	*Item response*
1. Formulation	Define construct	Define the basis for the item in terms of its content and cognitive behaviour related to construct
2. Explication	Test	Item
3. Validation	Evidence bearing on the interpretation and use of test scores for a specific purpose	Evidence bearing on the interpretation and use of an item response with other item responses in creating a test score that can be validly interpreted or used.

Thus, as good MC items are brought about through hard work and by taking different aspects into account, I propose to discuss the stages of item development in terms of both theoretically and practically oriented considerations of MC item construction as one part of the validation process.

2.6.1 Advice given to creators of multiple-choice test items

What does a good and transparent MC item look like? Experts in the field give some general pieces of advice for the construction of efficient and useful MC tasks. The challenges that are mentioned and consequently the advice that is given concern the nature of the stem, the key option and the distractors.

The stem ought to consist of a complete idea, preferably in the question-format - not of an incomplete sentence (Haladyna 2004: 68). The question should be expressed as concisely, simply, clearly and accurately as possible (Ebel & Frisbie 1991: 170) – the examinee must know what is being asked. The word NOT should be avoided, as it implies asking for an incorrect answer (Ebel & Frisbie 1991; Haladyna 2004).

As for the options, the most important principle is that the key option needs to be unambiguously correct and the distractors need to be unambiguously wrong. Even if, as Alderson et al. (1995: 47) say, this seems obvious, it is quite possible, especially in reading or listening comprehension tasks and in inferencing questions in particular, to write an answer that many colleagues would disagree with.

Not only should the distractors be clearly wrong, they should also represent plausible misinterpretations, attracting weaker test test-takers. (Cf. Linn & Miller 2005: 203; Haladyna 2004: 69,120). A plausible distractor will look like a right answer to those who lack the targeted knowledge or ability. As Alderson (1995: 48) puts it, each wrong alternative should be attractive to at least some of the students. Because if an alternative is never chosen, then it is wasting everyone's time and might as well not be there. This is logical as the purpose of a distractor in a MC item is to discriminate between those test-takers who have command of a specific body of knowledge and those who do not (Ebel & Frisbie 1991: 167).

As Haladyna (2004: 120) points out, writing plausible distractors comes through hard work and is the most difficult part of multiple-choice item writing. Some advice is given as to the way of creating plausible distractors. Ebel & Frisbie (1991: 167 ff) point out that plausible distractors should consist of responses that will enhance the clarity and efficiency of the item without providing irrelevant clues that lead the uninformed to the correct response. As for the types of responses that could be included, they suggest true statements that do not correctly answer the question presented in the stem, familiar expressions, phrases that have been used in common parlance or things associated with terms used in the question. An added advice for creating plausible distractors is to use students' most common errors (Cf. Linn & Miller 2005: 204; Haladyna 2004: 120), misunderstandings or carelessness and of inserting words with verbal associations to the stem.

Regarding the surface structure of the options, Ebel and Frisbie (1991) underline that the distractors should be parallel (similar) in grammatical structure, type of content, length as well as complexity. They point out that the choices should be clear and they consequently advice against including obscure distractors and too long or complex options. Distractors that are absurd or highly implausible should be avoided, as they will contribute little or none to the effectiveness of a test item. The key options should not be more general or more inclusive, or more careful, or more detailed or longer than the distractors. Haladyna (2004: 111, 118) adds the tips of avoiding the use of negatives in the options as well as specific determiners such as *always*, *never*, *totally*, *absolutely* and *completely*. In case specific determiners are needed in the right answer, they should also be included in the distractors (Cf. Mendelsohn & Rubin 1995: 43).

Linn and Miller (2005: 204) as well as Haladyna (2004: 103) put forth a word of warning: distractors should not be made into trick questions that mislead knowledgeable test-takers. They also caution against the strategy of inserting a negation in a correct answer to make it a distractor.

The quality of the options is related to the number of distractors worth including in an item. The number of options for test items is a matter of considerable debate. There is no magic number of alternatives to use in a MC item (Linn & Miller 2005: 202). As testing time is nearly always limited, there is a trade-off between the number of items and the number of choices per item. According to Linn & Miller (2005: 202; cf. Haladyna 2004: 112) three good distractors can be created, a fourth one being difficult to device. Each distractor should be selected by at least some of the test-takers. Ebel and Frisbie (1991) point out that if good distractors are available, the item is likely to be the more highly discriminating, the larger the number of alternatives is. In practice, however, more distractors usually mean that there are weak ones among them. According to Ebel and Frisbie (1991) it is quite possible to write a good multiple-choice test item with only one distractor (1991: 174). There seems to be a slight advantage of having more options per item, but only if each distractor is doing its job. As Haladyna (2004) points out, one criticism against the use of fewer options is that guessing plays a greater role in determining a test-taker's score. The advice for reducing the negative effects of guessing is to include more test items.

Rantanen (2003: 60) makes two conclusions on the number of options for a MC item based on theoretical and empirical findings. First, the number of options should be kept relatively low. Second, there cannot be claimed to be any ideal number of options, because of factors affecting the assessment circumstances: whether or not the options are created with the help of item analysis; where on the ability scale the most precise measurement is wanted and, how much time the processing of the options demand.

In a meta-analysis of research on the optimal number of options for a MC item, Rodriguez (2005: 3-11) agrees with Haladyna, Downing and Rodriguez (2002) in that we should use as many plausible distractors as feasible, but adds that in most cases, only three options is feasible: one correct option and two distractors. Rodriguez (2005) bases his conclusion on the fact that the use of more options does little to improve item and test score statistics and typically results in implausible distractors. He further points out the practical advantages of three options:

1) Less time is needed to prepare two plausible distractors;
2) More three-option items can be administered per unit of time, which may improve content coverage;
3) The additional high-quality items should improve test score reliability, providing additional validity-related evidence regarding the consistency of scores and score meaningfulness and usability;

4) More options result in exposing additional aspects of the domain to students, possibly increasing the provision of context clues to other questions (Rodriguez 2005:11).

However, Rodriguez (2005) would like to see more work done on the role of more effective plausible distractors, in search of information that increases our understanding of MC items so that the ability to measure student achievement would be improved[29].

In a test of listening comprehension, it seems reasonable not to have more than four options, since it puts weight on the reading comprehension ability, away from listening comprehension proper, especially in cases where the questions are in the target language.

Another issue related to the test rubric and the quality of the items is whether the test-takers should be allowed to see the item stem and options before listening to the text passage. The effect of different conditions was studied by Yanagawa & Green (2008: 107-122). They explored the differences in results in MC listening test performance between three different formats: the test-takers are either allowed to preview both the question stem and the answer options prior to listening, or just the answer options, or just the question stems. The version where only the options are previewed produces significantly fewer correct answers than the other versions. However, in case the question stem is pre-viewed, no additional benefit seems to come out of pre-viewing the options as well.

The researchers suggest that previewing the answer options may encourage test-takers to fall back on a lexical matching strategy. This can be self-defeating as reliance on the cues in the options may distract the test-takers from attempting to understand the broader relationship between the question and the text, relying instead on basic processes of recognition.

Another tendency seems to be that test-takers at lower levels suffer most in the absence of a question, as they may be less able to build a meaningful representation of a situation from the input without the support of the question. Changes in the test format on these variables seem to have less effect on test-takers at higher levels of proficiency. The researchers point out that even if the format with option preview only is likely to be more discriminating, a format that encourages lexical matching strategies must have limited validity for the testing of comprehension skills.

With reference to the limitations of their study, the researchers suggest that it might prove informative to combine test score outcomes with test taker reports on strategy use to further explore their hypotheses about why differences in scores emerge, i.e. evidence from the test-takers might provide valuable corroboration.

[29] On the discussion list of ILTA in early 2010 the question of the ideal number of options persists as a topic of heated discussions, with subjective and objective arguments presented defending different opinions.

Yet another variable in the test-taking situation is found with the number of times a listening passage is heard, reflecting interplay between cognitive and contextual factors (Geranpayeh & Taylor 2008: 2-5). It can be posited that if the text is played only once the listening situation would reflect more authenticity, as in a real-world context a piece of spoken text that has to be caught and interpreted is rarely heard more than once. The advantage for the practical test-taking conditions is that it permits more texts and types of listening activity to be sampled, as a larger number of test items can be included, thus yielding more response data. This would probably be a benefit from the point of view of both the validity and the reliability of the test results. However, the listening input would probably – at least for lower proficiency levels – have to be designed so that some internal repetition is present. Moreover, the items would, in the name of fairness to the test-takers, have to focus on explicit and easily accessible information.

The justifications for playing a listening text twice include the fact that the listening test context has an inherent artificiality to it, as most of the paralinguistic and contextual support features of a natural (TLU) listening situation are lacking. In a test situation there is seldom any interactivity nor is there the time needed to adapt to speaker accents and speaking patterns. The test situation lacks natural 'second chances'. Therefore, since a test is a test and not a natural or authentic TLU situation, a second listening would compensate for these lacking features. Moreover, being allowed to listen to a spoken text twice minimizes the risks of disturbances and lowers the anxiety of the test-takers, because they know that a second chance will be provided if the message is missed the first time.

In the current context of the Finnish Matriculation Examination, it is clear that tradition plays a role (factor also mentioned by Geranpayeh and Taylor 2008: 2-5) – the test specifications, or the examination traditions have an important influence on the test-taking conditions. The test-takers and their teachers are familiar with certain test-taking procedures, and it takes time to adjust to items that demand new ways of processing[30]. The specifications are related to the National Curriculum and therefore change rather slowly. This gives the involved parties – mainly students and teachers - time to adjust to new situations and demands.

Clearly the test-takers' proficiency level and their expectations on the test are also factors to consider. Today the test of foreign language listening ability in the Finnish Matriculation Examination, in the majority of cases, includes both items where the spoken text is listened to twice and test items that are to be answered after only one listening. This is related to considerations of validity: together with the inclusion of spoken texts of different nature, the two different conditions represent variables that contribute to a larger coverage of the construct at stake.

The general conclusion that can be made is that the MC item can be used for measuring listening comprehension abilities, even though these items are by

[30] Oxford et al.. (2004: 12) also stresses the influence of task familiarity and its effect on the processing load (see present chapter 1.5.2).

their nature complicated to construct. Care should be taken in the construction phase in order to make items clear and transparent, and not opaque and obscure, in which case we may have a situation where the test measures other things than listening comprehension and causes construct-irrelevant variance in the test scores.

2.6.2 Reviewing the multiple-choice items

As creating test tasks is a complex skill, and there are many ways they can go wrong, it is necessary to pre-test the items on a sample of test-takers similar to the target population, and then subject the results to item analysis (Buck 2001:148). Haladyna (2004: 183) recommends several reviews of a test as both research and experience have shown that many multiple-choice items are flawed in some way. He maintains that the time invested in these reviews will reap rewards in a direct way: the more polish we apply to items, the better the items become.

According to Bachman (2004: 120) we need to pre-test, or to try out our tests before they are used to make decisions since in this way we can analyse the scores and use the results of these analyses to make changes to improve the potential usefulness of the test. For large-scale tests, Bachman (2004) emphasizes, this generally involves trying the test out with a large number of individuals who are very similar in their characteristics to the individuals for whom the test is intended. Haladyna (1994: 127) refers to Messick (1989) in making a point about the value of reviews. Any factor contributing to the increased or reduced difficulty of the test or lack of discrimination that is external to the test content is a form of bias. It contaminates the inferences we require from test results.

Haladyna (2004: 190-201) presents a summary of MC item review activities that provide important pieces of validity evidence supporting both the validity of test score interpretation and uses and the validity of item response interpretations and uses (see Table 11 below).

TABLE 11 Item review activities (from Haladyna 2004: 201)

Item Review Activities	
1.Item-writing review	Checks items against guidelines for violations
2. Cognitive demand review	Checks item to see if its elicits the cognitive process intended
3. Content review	Checks for accuracy of content classification
4. Editorial review	Checks items for clarity and any grammar, spelling, punctuation or capitalization errors
5. Sensitivity and fairness review	Checks items for stereotyping of persons or insensitive use of language
6. Key check	Checks items for accuracy of correct answer
7.Answer justification	Listens to test takers alternative explanations for their choices and

	gives them credit when justified
8. Think-aloud	During the field test, subject each item to a round-table discussion by test takers. The results should inform test developers about the quality of the item for its intended content and cognitive process. Think-aloud is also a good research method.

Every item should be subjected to a review to see whether it was properly written (Activity 1). The items should further be considered from the point of view of the kind of behavior each item demands of the test-takers (Activity 2). According to Haladyna (2004) the central issue in content review is relevance. An expert (panel) should ensure that each item is relevant to the domain of content being tested and is properly identified as to this content. Content is believed to be definable in terms of a domain of knowledge and each test should be a representative sample of the total domain of knowledge (Activity 3). If there are many errors in the test, the test-takers are likely to think that the test falls short in the more important qualitative areas. This editorial review (Activity 4) concerns the face validity of a test. For the sensitivity review (Activity 5), Haladyna (2004: 193) refers to *ETS* (2003) and their recommendations, including the importance of treating people with respect, implying a balancing of the representation of groups and types of people, minimizing construct-irrelevant knowledge, avoiding controversial material, using appropriate terminology and avoiding stereotypes. The key should be checked (Activity 6) because in a superficially or casually constructed item there is the possibility that either there is no right answer, that there is a correct answer, but not the one intended, or that there are two or more correct answers. Answer justification (Activity 7) is a systematic study of correct answers from the standpoint of those who are going to or have taken the test – the answer justification provides the test developer useful information about how well the item works and complements the statistical analysis of item performance. For think-aloud (Activity 8), students are encouraged to talk about their approach to answering an item, while the administrator takes notes or audio- /videotapes the session. The value of think-aloud should be considered both in the setting of research and of item response validation. In research settings, the nature of the cognitive processes elicited by various item formats can be established, and the test specialists seeing the link between think-aloud and construct validity have recommended this practice. The outcomes of think-aloud in testing programs provide validity evidence concerning both content and cognitive behavior.

When implementing test items to a sample of test-takers similar to the target population at pre-testing, we can obtain essential information on the nature and functioning of the test items. The item performance patterns can be studied through item analysis by establishing item difficulty and item discrimination of each item, as well as the response patterns for each option. The results of an item analysis can be useful in identifying faulty items and can provide information about student misconceptions and topics that need additional work (Linn and Miller 2005: 350). Haladyna (2004: 203) explains that every response has a response pattern and some patterns are desirable and other patterns are undesirable. He says that the study of item responses provides a primary type of validity evidence bearing on the quality of test items. Item responses should follow patterns that conform to our idea about what the item measures and how examinees with varying degrees of knowledge or ability should encounter these items.

The quantitative item review activities are conducted *a posteriori* after a test or a pre-test has been administered. In the following, the basic principles of the Rasch analysis are treated.

2.6.3 Quantitative item validation

Assessment specialists have admitted to limitations of classical item analysis (classical test theory) in exploring the quality of test item use. Particularly in large-scale contexts classical item analysis has proved to be inadequate (Bachman 2004: 139, cf. Linn & Miller 2005: 358-360). In the first place, the limitations concern the fact that item statistics are sample-based and descriptive. This implies that it becomes difficult to compare items or test-takers from one test-round to another. Second, item statistics following a classical analysis does not take into consideration the connection between the difficulty of an item and the ability of a particular test-taker. Moreover, the test scores only tell us how well a test-taker did on average on a particular set of items. All items are treated as equal contributors to the total score.

Because of these shortcomings, measurement specialists have developed other models for relating a test-taker's performance on a given test to his or her level of ability. IRT rests on the assumption that the test-taker's performance on a given item is determined by 1) the test-taker's ability and 2) the characteristics of the item. One of the essential concepts and hypotheses behind the IRT is the latent trait theory, where the latent traits are conceived of as characteristics or attributes which account for consistencies in the individual's responses to items (Baker 1997: 19). There is thus a relationship between the observable performance on test items, and the unobservable abilities assumed to underlie this performance.

The advantages of IRT-models are first of all the fact that item parameter estimates are independent of the group of examinees used. Second, test-taker ability estimates are independent of the set of test items used (Bachman 2004: 142). IRT is thus probabilistic and inferential: both people and items are viewed as samples drawn from a larger population. IRT focuses on the pattern of item responses and the person and item measures are reported on the same scale. The person's ability measure is defined as the point on the scale where he or she is fifty percent likely to either succeed or fail.

Certain assumptions inherent in the response models relate to test unidimensionality and local statistical independence (Baker 1997: 29). Unidimensionality refers to the assumption that the item measures a single ability or trait and that the performance on the test should not be influenced to any important degree by other factors. Local independence implies that the probability of a person answering any one item correctly is not affected by information regarding that person's success on any other items. Therefore the content of one item must not provide any clues to the answer of another.

The Rasch model[31], used in the present study, is similar or equivalent to the one-parameter IRT model. In the Rasch model, the probability of success is assumed not to be affected by the possibility of guessing (Baker 1997: 27). The Rasch model differs from the two- and three- parameter IRT models in that it takes account of only one parameter (difficulty) while a two-parameter model also includes a discrimination parameter, and the three-parameter model also a guessing parameter (Bachman 2004: 141-142). While the one- and two-parameter models can be used with a sample of one or two hundred test-takers respectively, the three-parameter model requires a set of a thousand test-takers (Alderson et al. 1995: 91). According to Baker (1997: 61) a number of authors advocate the use of the Rasch model in preference to the more expensive and computationally more cumbersome procedures based on the two- and three-parameter models. One argument is the fact that the assumptions made by the Rasch model are in fact those upon which most tests are already based. Baker (1997: 61) quotes Pollitt (1979)[32] explaining that whenever a test is used to provide a score for each person, unidimensionality is implicitly assumed; and whenever this score is simply the number of items correct, equal discrimination and hence the Rasch model are similarly assumed. Baker (ibid) thus concludes that the Rasch model shares its basic assumptions with the traditional approach, but makes these assumptions explicitly rather than implicitly.

The use of IRT just as the use of classical item analysis, does not change the fact that statistics alone cannot determine which items on a test are good and which are bad, but statistics can be used to identify items that are worth a particularly close look (Livingston, in Downing & Haladyna 2006: 423) and in the final analysis, the worth of an achievement test item must be based on logical rather than statistical considerations (Linn & Miller 2005: 360). Item response models can thus be useful in test development, detection of biased items, score reporting, equating test forms and levels, item banking and other applications (Hambleton & Murrey 1983, in Baker 1997: 57). The models within the IRT are used extensively in practice in many test development projects, for item banking, for calibrating items or equating tests (Alderson et al. 1995: 92; Bachman 2004: 137ff) and at different stages in the validation process (Luoma 2001), for tests of listening as well as for tests of other skills.

[31] The class of models is named after Georg Rasch, a Danish mathematician and statistician who developed item response models and described them in his book *Probabilistic Models For Some Intelligence And Attainment Test* (Copenhagen: Danmarks Paedogogiske Institut. 1960). *B. D.* Wright was the leader and catalyst for most of the Rasch model research in the US throughout 1970s. His presentation at the ETS Invitational Conference on Testing Problems served as a major stimulus for work in IRT, especially with the Rasch model. In 1979, Wright and Stone described the theory underlying the Rasch model, and many promising applications (Wright BD and Stone MH (1979). *Best test design*. Chicago: MESA Press.) These were further developed in 1982 by Wright and Masters (Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.).

[32] Pollitt, A. 1979. Item banking. *Issues in Educational assessment*. SED Occasional Papers, London: HMSO.

In research contexts, the IRT or the Rasch model for investigating tests of listening comprehension are reported to have been used in studies by for example de Jong & Glas (1987), McNamara (1991), Buck (1994) and Brindley (1997). De Jong & Glas (1987) investigated the construct validity of L2 listening comprehension tests by means of evidence obtained from native speaker and target population data. They found that the measure of fit of items in the analysis of item responses from the target population paralleled native speakers' results on the items. Another result concerned the underlying trait of tests for L2 listening comprehension: literal understanding is preferable to interpretative understanding. McNamara (1991) examined the role of IRT in determining the validity of the listening subtest of the Occupational English Test for health professionals. By means of two statistical tests McNamara investigated whether it is possible to construct a single measurement dimension of listening ability from data from two different test parts, one with more abstract and one with more concrete content. The results indicate that it is possible to construct a single dimension using items from the test, with the items representing different levels and areas of ability. In McNamara's view, IRT proves to be useful in investigating the content and construct validity of language tests. Buck (1994) used the verbal report methodology to contrast the assumption of unidimensionality made by IRT models with the multidimensional nature of listening comprehension. His results indicate that the skills behind successful item performance vary and differ from one test-taker to another. The conclusion made is that the unidimensionality assumption is violated.

Brindley (1997) set out to explore the possibility of defining which listening skills are tapped by particular items, and to compare expected item difficulty with item difficulty evidenced by item analysis. His study showed that experienced language teachers have difficulties determining which specific skills are assessed by an item, and perceiving which items are more or less difficult for test-takers.

In this study, the software *WINSTEPS (*3.69.1), developed by Mike Linacre, is used to construct Rasch measurement from the responses of one particular set of persons to a particular set of items. It is important to point out that the quantitative analysis in this study has a focus on the individual items and not on the "test" used in this research context, which does not consist of an entire test but items drawn from a test administered for the Matriculation Examination. Thus the ability estimate given to the test-takers on the basis of these seventeen items cannot be used as a reliable indication of the test-takers' listening comprehension ability. In the present study, the Rasch analysis is therefore used with the objective of giving support for the qualitative analysis of the introspective responses.

2.6.4 Distractor evaluation

The development of good distractors will help the item perform as intended and the study of distractors is necessary for sound item and test development (Haladyna 1994: 19, 153). As with test items in their entirety, distractors that fail to perform should be revised, replaced or removed. The principle is that a distractor should appeal to low scorers – those who do not have the measured skill – whereas high scorers should avoid the distractors. If there are distractors that are never chosen, they are probably too implausible for any test-taker to select them.

Haladyna (2004: 209, 229) presents what he calls the most fundamental tabular methods: the mean of those choosing the correct answer and the mean of those choosing any incorrect answer. These provide clear summaries of response patterns. According to him, however, graphical methods are easier to understand and interpret. A graphical procedure – the trace line – (Haladyna, 1994: 153; 2004: 210-211) is a depiction of item performance as a function of total performance. An effectively performing item contains a trace line that is monotonically increasing – nonmonotonic trace lines should be viewed as undesirable. A trace line may show that a distractor fails to discriminate among the score groups – visible as a flat trace line. Haladyna (1994: 159) maintains that the trace line appears to offer the best, most sensitive and revealing picture of option performance, because it graphically illustrates patterns that can be easily understood, at the same time providing statistical evidence.

We can conclude, with Haladyna (2004), that all testing programs would benefit by adopting guidelines and studying item response patterns. Doing item response studies and taking appropriate action is another primary source of validity evidence, one that bears on item quality.

2.7 Validation in the context of the present research study

The decisions on item selection should be based on multiple sources of information, both qualitative and quantitative: content considerations with item statistics to supplement it (Bachman 2004: 137). Another important source comes from the analyses of the test-takers' processing of individual items.

Three main sources of evidence are within the scope of our validation study. First, the item contents (or context, using Weirs concept) needs to be analysed, both in relation to the construct and to the internal structure of the pool of items. To this adds the general information on the item construction stage provided by two of the responsible item developers for the original test items. [33]

Second, a Rasch analysis as well as a distractor evaluation is undertaken in order to obtain a quantitative basis for the content and internal test analysis. However, this information is worthless without an added qualitative analysis of the reasons behind these values and results:

> If one attempts to sidestep the most important part of test behaviour, which is what happens between item administration and item response, then one will find no clarity in tables of correlation coefficients (Borsboom et al. 2004: 1068).

[33] This information also serves as background to the construct for the original test and is presented in chapter 4.2.

Thus in this study the main source of evidence is that of test takers' response patterns and processes.

In order to provide as much evidence for and against the use of the listening test items under evaluation as is possible within my research framework, I propose to triangulate the validation study by applying three different types of sources: a qualitative item content analysis and a quantitative item analysis added to the main focus on an investigation of the test-taking processes using the introspection method. In the following chapter, I will describe and justify the methodology of collecting information by means of short written introspection.

3 INTRODUCTION IN THEORY AND PRACTICE IN THE ANALYSIS OF A MULTIPLE-CHOICE TEST OF LISTENING COMPREHENSION

In this study the main tool for analysis of the listening test items is the method of introspection. In the following, I will present the theoretical background of the method and quote some relevant studies in the field.

3.1 Theory: Why introspection?

In an attempt to pin down the more or less conscious processes and strategies activated in a test situation, the visible "product" of the test task - i.e. the answer to a question or the choice of options - does not give anything like a complete picture of what has actually happened inside the test-taker. Consequently, in a triangulation of the sources of evidence for the validity of a test and the quality of individual test tasks or items, introspection[34], a verbal report on the processes and strategies experienced by individual test-takers, is a valuable tool. Researchers like Bachman (2004) or Haladyna (2004), who do not themselves primarily focus on qualitative verbal report-based research methods, strongly recommend the use of think-aloud/verbal protocol analysis as a method of collecting information on test-takers' processes in the test review or tryout phase.

A general observation seems to be that it is difficult or even next to impossible for a native or an outside observer to fully predict the nature of cognitive behavior (Yepes 2001: 16), the processing demands (Brindley 1997:80), the comprehension schemata (Ross 1997: 216) and the cognitive constraints (Goh 2000: 56) experienced by the non-native speaker at particular tasks or items in deriving meaning from a spoken text in a foreign language. Through introspective analysis "*a new avenue to understanding the validity of item responses is poten-*

[34] The term *introspection* is used here and in the present study as a general term covering all types of verbal reports.

tially opened" (Ross 1997: 216). The introspective reports testees are asked to provide help to identify the processing demands, determine if the item (type) is testing what it purports to test (Yepes 2001: 16) and gain evidence on why an element of an assessment instrument is not behaving as anticipated (Green 1998: 34). According to Cohen (1998a: 38) the major purpose for using verbal report protocols is probably to reveal in detail what information is attended to while performing tasks – information that is otherwise lost for the investigator.

Green (1998: 34) states that quantitative methods allow us to identify particular items, item types or materials that are not functioning well but they do not provide us with direct information on the nature of the problem. At the stage of collecting validity evidence for an intended test use, the quantitative item analysis provides information the test developer can use in considering which items to reject, whether revisions to items are required, and what those revisions might be. The qualitative data obtained by introspection largely supplements the information provided by formal task analysis (Green 1998: 34, Ross 1997: 219). Cohen (1998: 39) points out that verbal reporting is not seen as a replacement of other means of research but rather as a complement to them, as all research measures have their potential strengths and weaknesses

3.1.1 Cognitive justification for introspection

The theoretical justification for the use of this method rests on the assumption that human cognition resembles an information processing system. The listening comprehension process is rarely entirely automatic, especially not in a L2 language, where an important part of the cognitive activities has to be controlled. The process demands such an intensive concentration on the heard "input" that a substantial portion of it remains in memory. Introspective reporting would therefore be enabled by the verbalization of the traces in memory (Yi'an 1998). Ericsson & Simon (1987: 25) present a framework for studying thinking, where verbal reports by subjects are seen as one of many types of observations that provide data on subjects' cognitive processes. A cognitive process is seen as a sequence of internal states successively transformed by a series of information processes. Within the framework of the information processing model, it is assumed that information recently acquired by the central processor is kept in short-term memory, and is directly accessible for further processing (e.g. for producing verbal reports) whereas information from long-term memory must first be retrieved (transferred to STM) before it can be reported. Whereas the neurological origin of cognitive processes may not be available for introspection, the cognitive events themselves can often be made available through verbal reporting. The directness of introspection gives it a character not found in any other means of investigating psychological phenomena (Cohen 1998a: 39).

3.1.2 The typology of verbal protocols

The types of introspection or retrospection collected through verbal reports or think-aloud protocols at a testing event vary a great deal. Faerch and Kasper

(1987), Cohen (1998) and Green (1998) give slightly different categorisations for the criteria related to the verbal report methodology, reflecting the principal methodological variables. There are, nevertheless, some common dimensions. One criterion concerns whether or not the cognitive information expected of the reports is related to a specific action or whether it is more general by nature. Cohen (1998a: 13) separates three types of verbal protocols: 1) self-reports that consist of general statements versus 2) self-observations and 3) self-revelations both of which are limited to certain events. Related to this, Faerch and Kasper (1987) consider as one criterion the object of introspection – whether it is cognitive, affective or social by nature.

The temporal dimension is essential – i.e. the distance between the event and the verbalization in cases where the reports are related to a specific action. Faerch and Kasper (1987: 15) speak about a continuum, with three points of reference:

1) Simultaneous = concurrent (or introspective within 20 seconds of the event according to Cohen (1998a))
2) Immediately consecutive/retrospective
3) Delayed consecutive/retrospective.

Green (1998; cf. Cohen 1998a: 13; 14) points to the clear favor of using concurrent reports, whenever possible in the language testing context, since they are less susceptible to influences from such unwanted variables as the filtering of information or "tidying up" (indicating what is believed to be required by the researcher, omitting information or giving of a false impression of completeness or coherence).

Another important dimension concerns the role of the experimenter – what kind of interaction is there between the participant and the experimenter: is the test-taker asked questions (in mediated verbalization) or is s/he prompted only if s/he pauses for a period of time (in non-mediated verbalization) (Cf. Faerch & Kasper 1987: 10-20; Green 1998: 4-7)? Another feature of the elicitation procedure is the data collection instrument: what degree of structure is imposed? The instruments can range from highly structured rating scales (as the Likert scale) or multiple-choice questionnaires to the least structured diaries or continuous think-aloud verbalizations (Faerch & Kasper 1987: 16-17).

One further criterion mentioned by Faerch and Kasper (1987: 15) is the possible training of the participants. Especially with introspective think-aloud verbalizations, it is said that a training phase is needed to get participants familiarized with the way they are expected to react (Green 1998: 16) or, in some cases, to raise their awareness towards the specific trait the experimenter/researcher is interested in (Cavalcanti, in Faerch and Kasper 1987: 239; Cohen 1998a: 16). Cohen (1998a: 37) points out the drawbacks of training respondents to use certain terms in their responses, because it may distort the data in cases where respondents are meant to supply their own accounts of their cognitive processes.

Several researchers discuss the issue of how prepared or how "naive" the participants should be and how structured an instrument should be. Allan

(1992, 1995, quoted in Alderson 2000:334) discovered that many students were not sufficiently verbal and found it difficult to report their thought processes. To overcome this, he attempted to use a checklist of predicted skills or strategies but found a) that the categories were unclear to students and b) that using the checklist risked skewing responses towards those that the checklist writer had thought of. Allan thus questions the validity of checklists, and advocates careful construction and piloting whenever this method is used. Cohen (1998a: 28) talks about the advantages of not relying on structured instruments when attempting to identify test-takers' strategies. In his view, the response to structured instruments may be simplistic or contain only brief information about any one learning strategy. He also warns about ambiguities in predetermined questions and about their explicitness, in that they may by their nature motivate the respondents to select a certain response. Goh (2000: 56) stresses the value of "the learner's voice" – the importance of providing opportunities for learners to report in their own words - in order to gain insights into their understanding and their comprehension difficulties.

If there are no lists and thus no limitations to the kind of introspective responses that the participants will provide, the researcher may meet with problems at the stage of sorting out the information obtained from the responses, since they are likely to be quite varied. Cohen (1998a: 28) sees the risk of reducing the structure of the instrument in the increase of the volume of data. The data themselves are likely to be more highly individualized, which could prevent the researcher from determining general patterns. It is, therefore, highly recommendable to conduct pilot studies before the stage of categorization and summarization of the responses has been reached.

3.1.3 Limitations of the use of introspection

The advocates of the introspection method do not look uncritically at the method but want to raise the awareness of its limitations. One source of limitation concerns the type of linguistic processing that is available for introspection. Part of the linguistic processing remains automatic and thus cannot be introspected by a language user (Yi'an 1998: 25; cf. Cohen 1998a: 36). In fact, we have a greater chance to learn more about the processes of comprehension when they do not "flow comfortably", that is when normally automatic processes shift to slow and controlled processing, where understanding is difficult to come by, where interpretation is only partially achieved, or where an attempt to communicate results in misunderstanding (Brown 1995: 42; cf. Yepes 2001: 17). Buck (1990) noticed that listeners often reported a lack of lexical knowledge and concluded that this was probably due to the fact that lexis seems to be the only form of linguistic knowledge that is accessible to conscious knowledge.

Other limitations as to the use of the introspection method have been mentioned and words of caution put forth. Gathering data during test-taking is not an uncontroversial matter, since the data-gathering can interfere with the test-taking process being investigated (Alderson et al. 1995: 176; Cf. Green 1998: 10; Cohen 1998a: 36). However, Buck (1990: 193) noticed that the interview format did not have much influence on the comprehension process, apparently because it allowed listeners more time to process and reconsider their comprehension.

Three potential problems that may affect the validity of the verbal reports themselves are mentioned by Green (1997: 10): 1) incomplete reporting (test takers leave out descriptions of processes they use), 2) distorted reporting (inaccurate reporting) and 3) extraneous reporting (descriptions of processes that are not actually used). The results of verbal report tasks are likely to vary according to the type of instructions given, the characteristics of participants (more or less informative, variation in verbal skills, different terms to describe similar processes, same terms for different processes), the types of material used in collecting protocols and the nature of the data analysis (Cohen 1998a: 37). It is important to keep in mind that introspection only provides data and that these data are open to some degree of subjective interpretation (Ross 1997: 236).

3.2 Practice: Previous studies using the method of introspection

The think-aloud procedure has been used at listening comprehension tasks for instance by Buck (1991; 6 participants, 54 items), Ross (1997; 40 participants, 25 items), Young (1997; 18 participants, 3 audio texts), Vandergrift (1998; 4 participants), Yi'an (1998; 10 participants), Goh (2000; 23 participants), Yepes (2001; 30 participants) and Wagner (2006; 8 participants, 18 items) where learners or test-takers have explained orally - in some cases with the use of various oral prompts - what they are thinking about while solving a specific listening task or an item. Listening questionnaires and strategy checklists have been applied by Vandergrift (2005).

In 1997, Ross (1997: 218) made the observation that introspection, as a method of test construct validation, was still rare in language testing for logistical reasons. The following year, Yi'an (1998: 27) pointed out that there are no research findings available concerning the effect of the MC format on listening comprehension, following the mentalist approach. Green (1998: 3) maintains that verbal protocols are starting to play a vital role in the validation of assessment instruments and methods.

There are reports on a few studies where different types of verbal protocols have been used in investigating the processes activated at a language test situation. The four research topics discussed by Buck (1991) were: 1) how the test method influences the comprehension process, 2) – 3) if it is possible to write test questions assessing higher-level processing and the extent to which listeners are monitoring their developing interpretation and 4) if and how question preview influences comprehension and test performance. 54 short-answer test items were administered to six test-takers in order to examine how listening tests work, and how processes not normally accessible through quantitative research methods influence testee performance, test reliability and validity (Buck 1991: 85). Buck (1991) maintained that the results supported earlier findings that the methodology can provide valuable insights into many aspects of language processing and how the processing relates to test performance.

Ross (1997) set out to study the inferencing strategies used by second language listeners, looking for the basis on which a picture was selected to match an utterance. He also wanted to know what the interaction of the aural processing of the input with the listener's introspective account was. The study was based on 10 out of 25 items on a picture identification test, where 40 listeners gave introspective accounts immediately after inferencing the meaning of the input and selecting responses. Ross (1997: 236) concluded that introspection affords a potentially valuable supplement to item-response theory as a technique of data collection, as it helps us to understand how the processing stages reached in listening are instantiated to visual referents.

Yi'an (1998) wanted to find out how subjects employ linguistic and nonlinguistic knowledge in performing on MC listening comprehension tasks, what effect the format had on the subjects' performance, and how the immediate retrospection would apply to that particular type of study. The procedure included listening to the input twice, the first time to the whole text while completing the task, the second time section by section, immediately followed by introspection. The results and discussion were based on the data provided by 4 subjects (out of 4 in pilot study + 6 in main study). Yi'an concluded that the method worked well for the investigation of second language processing.

Among the studies focusing on tests of listening, one study is reported in an unpublished research paper by Yepes (2001). The researcher wanted to investigate if computer-based minitalks were more difficult than paper and pencil-based minitalks at a *TOEFL* test of listening comprehension, what factors affected item difficulty, and what the relationship between item difficulty and response types was. The type of method used included immediate introspective interview questions asking what the 30 test-takers were thinking and why they had selected a particular answer at each of the 25 items. The researcher found the introspection method helpful as a tool for identifying the processing demands and for determining what an item is really testing.

In his doctoral thesis, Wagner (2006) set out to investigate the role of the visual component for the listening test-taking context, looking at the interaction between test-takers, the listening text, the video and non-verbal information. The study consisted of several parts using different methods, and the third part of the study focused on how the test-takers reported utilizing non-verbal information in order to process spoken text. A total of 8 participants provided concurrent verbalizations while viewing three videotexts, concurrent and retrospective verbal reports while answering 18 comprehension questions and retrospective verbalizations after they had completed the test task. Wagner (2006) concluded that the evidence provided by these verbal report data -though incomplete - is informative and useful for examining the participants' thought processes, completing the information from the more quantitative data found in the other two parts in the study.

In a study of reading comprehension tests, Nevo (1989) purported to investigate whether it was possible to obtain feedback from respondents about their strategy use on an item-by-item basis. Nevo wanted to know if there is transfer of strategies from L1 to L2, how frequently strategies used in taking L1 and L2 MC reading comprehension test contributed to correct responses as opposed to incorrect ones, and if the stimulus format or response format of the test influences the selection of strategies. The instrument consisted of a MC test on four reading passages (two in Hebrew L1, two in French L2), open-ended questions concerning the respondents' evaluation of the test items, and a checklist of strategies for immediate introspective use following each of the reading passages or items, and at the end of the test, a retrospective questionnaire on more general strategies.

The conclusion on the usefulness of the method was that it is possible to get feedback from the subjects immediately after each item concerning the way in which they processed that item. A variety of data was obtained through the method, allowing the researchers to look into the minds of the respondents within seconds of their task-performance and to extract meaningful and authentic information, which they could not have collected by means of other instruments. Nevo (1989: 212) mentions a supplementary benefit of the verbal reports, namely that the task made the respondents aware of what they actually did and the strategies they used when taking a test.

Alderson (1990) who used both concurrent and retrospective self-observation with two test-takers in a pilot study has shown that introspection during a reading comprehension task was useful in identifying weaknesses in test items. Alderson (1990: 478) states that what a test of reading tests is not simply what its constructors say it tests, nor what a set of judges considers it to test but it must surely and crucially relate to what happens inside a test-taker's head when he or she responds to an item.

With the seventeen test-taking strategies selected among a range of strategies established by earlier studies, partly based on Nevo's (1989) MC strategy checklist, Anderson et al. (1991) studied what information the triangulation of data sources (test-taking strategies, item content and item performance) can give for construct validation of a reading comprehension test. The test consisting of 45 items was administered to 28 Spanish-speaking students of English. The researchers made the general conclusion that more than one source of data needs to be used in determining the success of test items.

In his doctoral thesis, Allan (1992, as reported in Alderson 2000: 331) gathered introspections to investigate strategies used to answer multiple-choice items and free-response questions on a *TOEFL* reading test, and concluded that MC items engaged strategies which focused more on the stem and alternatives, whereas free-response strategies centered more on the text passage and the student knowledge of the topic. In addition, he found that MC items engaged test-wiseness strategies.

Alderson (2000: 97) stresses that test-taker introspections show that the process of answering test questions varies from one reader to another. Thus an item may be measuring different skills for different persons. He points out the important validity issue of knowing what processes underlie responses that are considered correct.

Rupp et al. (2006) in a study investigating the differences between the response processes induced when responding to MC questions compared with processes drawn on when reading in non-testing contexts, used the verbal report method with ten participants who were given three reading passages with several MC questions. The participants in this study were prompted using a semi-structured interview format while responding to reading comprehension questions. The participants were also asked to think-aloud during responding to the MC questions.

Rupp et al. (2006: 457) summarize the think-aloud methodology as being one of the major tools used in qualitative studies investigating reading comprehension, since it provides an indirect view of a reader's mental processes which are unobservable during silent reading. In the present study I shall assume that the same holds true for listening comprehension. The type of verbal protocol tool used in the current investigation is described below in chapter 6.3.

4 THE BACKGROUND TO THE TESTING CONTEXT

In order to bring light on the current testing context and the original construct, I will briefly explain the language assessment situation in Finland. This is followed by a description of the original test of listening comprehension as it was used for the Matriculation Examination and of the context surrounding the test setting. Moreover, two experienced test constructors describe the test construction principles and procedures, information collected during an oral interview. This evidence on the item construction can also be taken as one step in the validation process, following the recommendation by Haladyna (2004: 1) and Anastasi (1986: 3).

4.1 Language assessment in Finland

A general tendency in Finland is not to assess language abilities “from above”, by means of national tests or assessments (Tarnanen et al. 2007: 384). The procedures of assessment and evaluation of Finnish pupils’ language skills are undertaken throughout their education almost exclusively by their own teachers. This evaluation is therefore heterogeneous and assessment procedures vary greatly, as a function of the teachers’ assessment skills, preferences and foci. The classroom-based assessment may consist of traditional tests but also for example of assessing oral or written presentations, individual or shared projects or the pupil’s activity in class. The basis for the evaluation procedures should naturally be the same as the basis for the teaching - the curriculum. However, even if the objectives for learning a foreign language, the central contents and evaluation criteria are described in the Finnish National Curricula in rather detailed manners, they do not take a stance on how the achievement of these objectives should be evaluated, and to what extent the different sub-skills should be assessed (Tarnanen et al. 2007: 384). Many teachers make use of the tests - as such or modified - that come with a textbook series for example, and they may use the tests both as diagnostic (formative) and as summative assessment tools.

An important change is emerging within the language examination systems also in educational contexts, as the bases for language curricula are altered to correspond to criterions for the different ability levels. This change enables more pedagogic, more transparent and more ethical assessment (Tarnanen et al. 2007: 384). The emphasis lies upon criterion-referenced assessment, where the student's/learner's performance is being compared with the criteria on the different levels, and not with a normative group performance. The foundation for this change consist of the *Common European Framework of Reference for Language Teaching, Learning and Assessment* (*CEFR*), which has helped to standardize the language didactic concepts and which is helping to create internationally comparable examinations throughout Europe and beyond.

These changes naturally put emphasis and demands on the knowledge and skills of teachers, who will have to - to a lesser or greater degree - adjust their assessment practices to the new standards. It is both an individual and a collective challenge, where individual and subjective practices and institutionalized traditions meet the new requirements.

There are, however, three external national and standardized language assessment systems in Finland. The Matriculation Examination is related to the school context and constitutes the school-leaving examination for upper-secondary pupils. In addition, there are two independent examination systems for which the focus is on the need of adults for the labour market, for study or immigration purposes. These are the National Language Examination in Swedish and Finnish (for Civil Servants) and the National Certificates of Language Proficiency in several languages (Finnish, Swedish, Sami, English, Spanish, Italian, French, German and Russian).[35]

The levels of the National Certificates of Language Proficiency are said to correspond to the *CEFR* levels[36], implying that the low levels 1 and 2 would correspond to the *CEFR* levels A1 and A2; the intermediate levels 3 and 4 would correspond to the *CEFR* levels B1 and B2, while the advanced levels 5 and 6 would correspond to the *CEFR* levels C1 and C2[37]. The levels for the two national examinations correspond to each other according to the following Table 12[38]:

[35] Many foreign international language examination systems also administer their tests locally in Finland, such as *TOEFL, IELTS for English; DALF, TCF* for French and *TestDaF* for German, among many others.

[36] See Appendix V.

[37] http://www.oph.fi/pageLast.asp?path=1,17629,3432,3435 2.10 2008

[38] http://www.oph.fi/pageLast.asp?path=1,17629,35053,35070 2.10 2008

TABLE 12 Comparison of the levels of the Finnish National Certificates of Language Proficiency and the National Language Examination (for Civil servants)

National Certificates of Language Proficiency (YLEINEN KIELITUTKINTO)	The National Language Examination (for Civil servants) (VALTIONHALLINNON KIELITUTKINTO)
Speaking and listening comprehension	**Oral skills**
• skill level 6 • skill levels 5 and 4 • skill level 3	• excellent skill • good skill • satisfactory skill
Writing and reading comprehension	**Written skills**
• skill level 6 • skill levels 5 and 4 • skill level 3	• excellent skill • good skill • satisfactory skill
Listening and reading comprehension	**Comprehension skills**
• skill level 6 • skill levels 5 and 4 • skill level 3	• excellent skill • good skill • satisfactory skill

4.1.1 The Finnish Matriculation Examination

The Matriculation Examination was first organized in Finland in 1852 as an entrance examination to Helsinki University, where the test-takers were expected to show sufficient evidence of an all-round education and knowledge of Latin. Nowadays, the purpose of the examination is to discover whether pupils have assimilated the knowledge and skills required by the curriculum for the upper secondary school and whether they have reached an adequate level of maturity in line with that school’s goals. Passing the Matriculation Examination in different subjects entitles the test-taker to continue his or her studies at university level. The examination is organized simultaneously in all Finnish upper secondary schools biannually, in spring and in autumn.

The Matriculation Examination Board is responsible for administering the examination, for its organization and execution. The Board issues guidelines on the contents, the organization and the assessment of the tests. [39] Because the Matriculation Examination is public and administered in the schools by the teachers, new tests have to be produced twice a year.

The numbers of test-takers for the different foreign languages in the Matriculation Examination in spring 2011 are as presented in Table 13 below and indicate the general proportion of the test-takers in the examination as well as of students learning these different languages at school[40]:

[39] www.ylioppilastutkinto.fi/english.html 11.3 2008

[40] Numbers provided in: www. ylioppilastutkinto.fi/Tilastoja/ KEVÄÄN_2011_YLIOPPILASTUTKINTO.pdf

TABLE 13 Numbers of test-takers in the Finnish Matriculation Examination for different foreign languages in spring 2011

Foreign Language	A (advanced level) Number of test-takers:	B (intermediate level) Number of test-takers:	Total number of test-takers:
English	23762	552	24314
French	338	1305[41]	1643
German	730	2039	2769
Italian	-	118	118
Latin	-	40	40
Portuguese	-	15	15
Russian	198	452	650
Spanish	44	1023	1067

The levels for the language tests in the Matriculation Exam are determined according to the levels created for the National Curriculum. They correspond to the CEFR levels, but contain descriptors for further sub-levels. The target levels demanded at the end of upper secondary school for a second or a foreign language studied for a varying number of years depending on if it is introduced at primary (so called A-language), secondary (B1 or B2- language) or upper-secondary school level (B3-language) are presented in Table 14 below. I will, in the following, rename the A-language "advanced" and B "intermediate", in order not to mix up the school levels with the CEFR levels.

TABLE 14 The target levels for the different skills for different L2

Language and amount of studies	Listening comprehension	Speaking	Reading comprehension	Writing
English (Advanced)	B2.1	B2.1	B2.1	B2.1
Second or Foreign Language (Advanced)	B1.1–B1.2	B1.1	B1.2	B1.1–B1.2
English (Intermediate B1)	B1.2	B1.2	B1.2	B1.2
English (Intermediate B2)	B1.1	B1.1	B1.1	B1.1
Foreign language (Intermediate B2)	**A2.2**	**A2.1–A2.2**	**A2.2–B1.1**	**A2.1–A2.2**
English (Intermediate B3)	B1.1	A2.2	B1.1	B1.1
Foreign language (Intermediate B3)	**A2.1–A2.2**	**A2.1**	**A2.1–A2.2**	**A1.3–A2.1**

As can be noticed in Table 14 above, the targeted levels differ both as a function of the subskill, the level and as well as the language, as it has been judged that higher target levels for English can be reached compared to any other language[42]. Apart from the second national languages (Finnish for Swedish-

[41] This figure can be compared with the number of test-takers for French as a FL in 2007 (N: 1979), and in 2002 (N: 3262) – illustrating the tendency of decrease.

[42] English is the most popular - the first - foreign language, a language everyone is supposed to manage to some degree, necessary for a peripheric country with an internationally completely unknown language, for international contacts, politics, business etc. This is evidenced by the large numbers of test-takers every year, see Table 13.

speakers and Swedish for Finnish-speakers) and English, the foreign languages for which there exist tests are French, German, Russian, Italian, Spanish, Latin and Sami.

A somewhat paradoxical situation arises when the different skills and proficiency levels are related to the criteria described in the CEFR, although the examination remains norm-referenced (described in 4.1.3). There is a clear controversy in the entire system, to the great confusion of the people working with the language examinations. It would seem meaningful to describe the test-takers' results in terms of their actual abilities such as they relate to the criteria, not according to traditional norm-referenced grades.

For the purpose of the current study and the analysis of the test of listening comprehension, the levels that are interesting are those for the foreign languages on the intermediate level (language B2 or B3). Students who have studied French from secondary school take the same exam as students who have started French in upper secondary school. For listening comprehension, the targeted average level is of A2.1-A2.2 (described below for the purpose of establishing content validity for the items under scrutiny), which corresponds to the level A2 on the CEFR scale. The French language examination under the supervision of the French Ministry of Education on a corresponding level would be DELF A2, or a test TCF on the level A2.

The levels that are linked to the CEFR were introduced into the revised national curriculum as late as 2005, so it was not yet valid at the time of the administration of the test under study, in 2002. However, the school systems, the number of lessons spent on studying the foreign languages, the teaching methods or the quality of the skills of the students have probably not changed dramatically over the two years, so it can be assumed that the target level in 2002 was fairly identical to the level described here.

4.1.2 The original testing context: the foreign language test

In the case of the test of listening comprehension of French as a Foreign Language at the Matriculation Examination, there are several factors we have to consider in "reconstructing" a construct that is not explicitly stated in the scarce information on the test provided by the Matriculation Examination Board[43]. One factor is the nature and the principles of the Matriculation Examination of Foreign Languages as a whole. As a result of having passed a Foreign Language Test, the participants get an overall score for their ability in that foreign language, where the scores for separate subskills are not specified. The Examination is a school-leaving exam, that is, based on the education provided by the

[43] What is described here is based on the situation at the time the particular test used for this study was administered, that is in spring 2002. Since then, both the National Core Curriculum, on which the Finnish Matriculation Examination (FME) is based, as well as the specifications for the test of foreign languages at the FME have been modified (as of 2005 for the curriculum, and 2006 for the FME). Among other more or less important changes, today the curriculum as well as the examination relate closer to the CEFR. See www.ylioppilastutkinto.fi; www.oph.fi/SubPage.asp?path=1,17627,1560.

upper secondary school. However, it is also a proficiency test, intended to give an overall score of the "general" language ability of the participants, no matter what school they have attended or what teachers they have had. It is a national test, with the National Curriculum as a base on which both the teaching and the test are expected to rest.

The entire examination of French as a Foreign Language in 2002 consisted of two non-equivalent parts:

- A test of listening comprehension, given approximately a month before the second part;
- A test of reading comprehension and written production consisting of
    1) a test of reading comprehension
    2) a test of vocabulary and grammar (typically a cloze test)
    3) a short essay.

The total maximum score is 299 points, with the test of listening comprehension giving a maximum of 90 points. The total score is then "translated" into an equivalent on a 6-graded scale that will be dependent on the general success of the participants who take the exam at a specific moment. The nationally defined grades (from improbatur (fail) to laudatur (excellent)) are given to the test-takers at one testing occasion in more or less following proportions: L: 5%, E: 5%, M: 20%, C: 24%, B: 20%, A: 11%, I: 5%

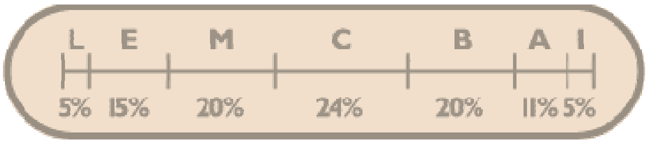


FIGURE 4 Proportion of different grades given in the Finnish Matriculation Examination

The relative shares of the grades differ somewhat in various tests and for various examination periods. The grades are then given on the final certificate together with the assessments of other subjects included in the exam. The grades serve, for instance, as one of the selection criteria for entrance into universities.

There is no explicit conceptual or theoretical framework given by the Matriculation Examination Board for the Foreign Language Test in the year 2002. A majority of the principles and rules provided concern the practical implementation and the operational construct of the tests. It is, of course, essential that all testing conditions are as similar and equal as possible, that the quality of the room provided for the test is as optimal as possible - especially for the test of listening - that the technical equipment necessary for the test is in order. In principle, there are no nationally defined compulsory courses for the foreign language studies begun at secondary school (approx. at the age of 14) or upper secondary school (at the age of 16), but the starting point for the construction of the test is the description of the goals in the National Curriculum for the total of eight L2 courses offered at upper secondary school. Having passed a minimum of three courses out of eight, the pupil is allowed to take part in the Foreign Language Test.

In the test of listening comprehension the participant is supposed to respond to the questions in the test booklet on the basis of the audio tape/recording. He or she is not to digress from the instructions stated both on the recording and in the test booklet.

In order to get at a possible conceptual listening construct on which the test is based we have to turn to the National Core Curriculum and to focus on what is said about the listening comprehension ability. First, general goals for the foreign language courses are given, some of which are related to the element of listening comprehension. The general goal of the curriculum defines the following objectives:

- The pupils will be able to communicate in the foreign language in various everyday situations.
- They can handle the language- specific communication strategies.
- They are able to develop their language ability.

For the intermediate (B3) level language, i.e. at the level comparable to the courses beginning at upper secondary school, specific goals are stated and include the following:

- The pupils understand the main content in everyday language spoken at normal pace;
- They master the central vocabulary and expressions as well as the basic structures;
- They master the main communication strategies typical of the target language and culture.

Thematic fields or content areas are also mentioned. These include: the human being and his/her environment and life situation; society; work, studies, spare time and hobbies; current events and the media; literature, music, art, film; geography; history; the ethic-religious values; science, economy and technique; environmental education and the nature; the home country, the surrounding world and internationalism. A point of departure for the selection of vocabulary and structures is their communicative value, their frequency and their usefulness within different fields and in different spoken and written situations. It is wished that the pupils be taught purposeful learning strategies. When learning the grammatical structures and rules, the pupils also develop their ability to infer on the level of vocabulary, sentences and texts. Moreover, in order to develop their communication ability and to give the pupils a natural pronunciation and a good accent, rhythm and intonation, they have to be provided frequent possibilities of listening to authentic recordings of the target language, including regional variants.

We can see that the conceptual construct has to be built upon a very broad but vague base. It is self-evident that the simple construct of L2 listening comprehension consists of the ability to understand the foreign, target language in its spoken mode. The content fields mentioned are of such a variety that they cover practically every thinkable theme. No language functions are stated, except for a general ability to communicate conveniently according to language- and culture-specific traditions. The "amount" of language that a pupil can master after approximately 300-350 hours of teaching sets its evident limitations – in reality, after having acquired the basic structure and functions of a language, there is not much time left for learning theme-specific vocabulary or functions, therefore only scratches on the surface will be possible. The concept of "everyday language at a normal pace" is rather vague, and it is up to the test constructors to define what that should cover.

A lot of the responsibility for selecting appropriate course contents for use in schools lies in the hands of textbook writers[44], who often also provide auditory material to be used alongside the textbook. The main responsibility lies, however, with the language teacher, who usually has the authority to select the textbook and other material that he or she wants to include in the language course. It is recommended that as much authentic material as possible is used, but it depends very much on the individual teacher what his or her ambitions are, how much time he or she is willing to spend outside class, looking for suitable extracts from TV, radio and the Internet, and how much time is spent in class listening to these authentic materials. Time is always limited, and teachers, in their liberty of choice, have to carry out a very rigid selection process in order to be able to treat even the compulsory contents of the courses. French clearly being a foreign language in Finland, people who want to learn it actually need to make a great effort to be able to listen to spoken specimens of the language. Even if the Internet opens new doors to all kinds of sources of language, in practice learners get to listen to French mostly in class. The amount of time spent listening to French is not large; as teaching includes all different pieces of language needed to build up a more or less communicative ability from the total beginning, within a maximum of approximately 350 hours of French, the amount of listening is by necessity scarce. This is naturally a pity, since the spoken mode of communication is its basic element, and since many students would like to learn more spoken communication. Listening comprehension is a prerequisite to speaking, and necessary for the ability to learn intonation and prosody. However, there are simply so many other elements to treat in class, that very little time is spent listening (and speaking). Ideally, there should be many French-speaking natives around as "resources" in schools, in order for the student to get the possibility of practicing authentic listening and speaking. As this is usually not the case, the consequences for testing are clear: there cannot be texts that are specifically authentic, with respect to specific characteristics related to the spoken language: hesitations, unclear speech, repetitions, false starts, etc. This would simply be unfair to the test-takers, since they have not had the opportunity to practice this type of text processing.

At a certain point, it becomes rational for the teacher to turn to previous language tests, to get hints on what type of knowledge and skills it pays off to include in the course – what specific details of grammar and what type of vocabulary have been focused by the test constructors during earlier tests? In

[44] See also Kauppinen et al. (2008) on the contents of FL textbooks for pupils on the 9th grade.

other words, there will be *washback* (or impact), i.e. the direct or indirect influence a test has on the teaching, and learning to the test. This might lead to a vicious circle: the test contents should be based on the teaching contents but, at the same time, the test contents are likely to influence the teaching content (illustrated in Figure 3).

For the test of listening ability, this also concerns the type of items included in the test. So far, the test of listening comprehension of French as a foreign language at the level B3 has mainly included MC items, with recent inclusions (after spring 2002) of short, open-ended questions, and sometimes true/false statements.

This relates to the selection of the type and themes of spoken texts, and to the types of test items that could be expected to be familiar to pupils at this level, and thus included in a test of listening comprehension. Since there is no explicit construct description that would provide the conceptual framework for the test, the test evaluators as well as the test developers have to rely on what there is at hand: the National Core Curriculum, textbooks in use, "common sense" and tradition. For the test developers this places great demands on their experience and their reliance on the tradition of test construction within the framework of the Matriculation Examinations as well as on their competence and creativity. What is the test constructors position and point of view? How does the test construction process look like in reality? These questions are closely linked to the validity of the test. Some important issues are presented in the next section as a result of an interview with two experienced constructors of the test of French as a foreign language. Their statements give us some information on the conceptual construct, but more, perhaps, on the operational construct for the current listening comprehension items.

4.2 The test constructors' point of view: background on the test context by two item writers & test developers45

The Matriculation Examination board gave permission to conduct an interview with the two representatives from the French Division in the Language Committee. I had prepared some general questions for the interview event, and met personally with both the two test developers at the same occasion. They provided detailed information on the test construction procedure. This information serves as more background for the task of trying to establish a construct to use as a basis for the validation procedure, at the same time representing one step in the validation procedure. The information collected by means of the interview is reported in the following.

[45] Open interview conducted by the researcher on the 24th of October, 2003; thus representing the conditions and circumstances at that time. I wrote down the responses given by the test constructors and let them read through my summary a week after the interview.

4.2.1 General information

The construction of a foreign language test for the Finnish Matriculation Examination begins 1 – 1.5 years before the test is to be administered. After a test included in the examination has been administered at the different schools, it becomes public. New test batteries therefore have to be produced twice a year. The French division, consisting of several people of which some are both item writers and censors, and some are either or, first meet to discuss how the different tasks associated with the construction of the entire test of French as a foreign language are to be divided between the members. Some members are specialized in test construction and they create – among other tasks - listening comprehension test suggestions that are, eventually, presented to the Language Committee (*kielivaliokunta*) of the Matriculation Examination Board, with members representing different foreign languages. Moreover, there are native speakers of the different languages, as consultants, as well as experts on language testing in general. People in the field, i.e. former or working teachers, are also consulted and asked to provide feedback.

The framework of the curriculum of foreign languages for upper secondary school is kept in mind, as well as the content of the foreign language textbooks. The constructors obtain information on which textbooks different teachers in different schools use when the test materials are returned from the field. This is necessary in order to construct the test on a suitable level both regarding the structures and the vocabulary covered in the foreign language courses. As the new curriculum (that has since come into effect in 2005) will follow the recommendations of the *CEFR*, the constructors will eventually be trained in implementing the principles of this framework.

4.2.2 The procedure of constructing a test of listening comprehension

The procedure of constructing a test of listening comprehension follows the same patterns from one occasion to another. First of all, possible text alternatives are chosen. These are mostly found in newspapers, magazines or on the Internet, that is, normally in the shape of written documents, occasionally auditory ones - even if the last ones can only be used for pupils who have studied the language as an advanced language, that is beginning in primary school (from the age of 10-12). Usually they are too demanding for the pupils on the intermediate language level (the target level for the present test items), beginning in secondary (at the age of 14) or upper secondary school (at the age of 16).

The test of listening comprehension as a whole will consist of at least three different texts, to get variation and to avoid bias. The procedure of selecting the spoken texts for tests on the intermediate level follows such principles as: topics should be familiar to young people, texts should be neutral, concrete and treat everyday topics, be "spoken language-like" with text types like interviews (the easiest and the most common), narrative texts or messages. On the other hand, texts that are avoided include those treating topics of too current an interest, texts in non-logical order, without a clear message/structure, issues that are too abstract as well as texts that are discriminating or morally judging, or linked to some kind of ideology.

There are several types of modifications that are undertaken in order to produce a suitable text. First of all, the text is shortened – often a great deal. It is simplified both with regard to the lexicon (more frequent words replace less frequent, argotic or technical ones) and the syntax (heavy constructions belonging to more formal written text are replaced). The language that is used in the final version of the text should be plausible but yet straightforward, therefore, spoken language connectors are added to original written texts while confusing redundancies are deleted from original spoken texts. Sometimes major modifications are made so that only the main ideas of the text are retained, changing its format completely, in practice creating new texts. By contrast, the scripts for the short communicative situations, representing one subtype of items, are written by the constructors themselves.

Certain general principles are followed in creating tasks or items. One concerns the length of a passage: if it is associated with two MC items (as it usually has been in the traditional format), it should not be more than one minute long. Apart from MC items, other item types are possible such as, open-ended, correct/false or not stated, cloze tests and résumés. (This last type is not used on the intermediate level, however.) The questions of the items should focus on the central, essential information in the text. The tasks are supposed to reflect the capacity of the test-takers to perceive the main ideas in the text. Tasks focusing details – numbers, for example – are necessary, too. However, there should not be more than two of them (per 30 items). The test as a whole should present a suitable range of texts and tasks.

The texts used as a base for the MC questions should include sufficient material for creating distractors. Three MC alternatives are preferred to four, since it is often difficult to create three plausible distractors. However, the two distractors should both be attractive. Moreover, there should be a higher total number of items, in order to minimize the random guessing factor. This principle also applies to true/false – items: if the T/F format is used, there should be at least 10 items of this type.

The questions and the options are all given in the written mode. The MC alternatives should ideally be short and have the same structure. There can be items where the test-takers are allowed to listen to the spoken text either twice or just once. The open-ended questions are given and answered in the test-taker's native language.

It is interesting to note that the principles above, mentioned by the test constructors, are indeed based on theoretical knowledge as well as on experience in test construction. They seem to illustrate the "optimal procedures" used by knowledgeable constructors, but with the compulsory restrictions set by the practical conditions surrounding the construction process.

When the texts have been modified and tasks have been created to accompany the texts, the propositions are first discussed within the French division of the board. The propositions are modified and revised, taking into account the opinions of the teacher representatives. After this, the whole is presented to the

Language Committee. The committee might object to the texts, and in that case they are sent back to the constructors to be revised, changed or replaced by other texts, and discussed again in the French division. The items are checked in detail to find possible problems, and native French speakers also scrutinize the test.

When the texts and the tasks have been approved of, the test booklets are printed and presented to the language committee for one more check. Then experienced native speakers - who are used to reading in a clear, sufficiently slow but natural way - read or act out the scripts for recording. Sometimes a speaker may deviate from the script on some detail, in order to make the text sound even more authentic. No background noise is added, though, as the test-takers are not used to it. The recorded material is then checked by the French division and by an independent exterior person. At this stage, the test as a whole is approved by the Matriculation Examination Board. There is always a complete supplementary test kept in reserve, in case some unpredictable problem with the main test should occur.

As the Language Committee treats all the foreign language tests, and naturally also all different sub-tests (listening comprehension, reading comprehension, cloze and written production) of the entire examination, there is influence from both the other languages of the Matriculation Examination and the other sub-tests on the test of listening comprehension. The other tests provide information on the language level used for the different languages, and present ideas on different types of items. According to the constructors, French as a foreign language at this level does not differ much from the other foreign languages as a test object.

4.2.3 Other comments on the listening comprehension test

In their evaluation of the Foreign Language Test as a whole, the constructors regretted the lack of one of the parts of the concept of communicative competence: the oral production skill. A combined test of listening comprehension and of the ability to speak would be a natural way of testing oral competence. However, this has been considered impossible for practical reasons.

The constructors both underlined the importance of testing listening in a separate sub-test, since it is an essential part of the communicative competence not possible to test in terms of other sub-tests. Moreover, for some test-takers the listening comprehension might be their preferred or strongest competence, giving them a chance to obtain reasonably good results in the L2 exam as a whole, even if they are not equally strong in written skills.

In discussing the outcome of the administered tests, the constructors were frustrated at the fact that it sometimes is impossible to foresee how a certain item works out. An item considered relatively easy can turn out to be difficult for the test-takers. An example of this was an open-ended question on identifying the two persons involved in a dialogue. The dialogue was felt to be very easy, with several hints included in the text with indications of the actors being on a train, one being the guard/ticket-collector, the other one the passenger. In their responses, many test-takers were, however, confused by the identity of the guard, not understanding that the scene took place on the train.

On the whole, though, the test constructors' point of view is that the reactions of the test-takers are reasonably predictable. This is clearly seen with MC items, where the choice patterns follow the intended effects of the distractors, i.e. weak test-takers choose the weak distractors, strong test-takers the strong ones.

4.3 The original test procedure of the subtest of French listening comprehension

In order to be able to compare and validate the current test procedure and analysis, I provide here a description of the original test procedure – as an indirect result of the construction process described above. This procedure will be referred to in the analysis that follows.

The original test of listening comprehension administered in spring, 2002 consisted of four parts, according to the following descriptions:

I. The first text consists of five passages making up a text and a thematic whole: the phenomenon where people buy old shops and offices in order to live in them. The first passage introduces the theme (as an argument) and the rest give supporting evidence in terms of six different people sharing their experiences. Eleven MC items with three options each including one key are created on the basis of this text. The participants listen to the entire text once, after which they listen to it a second time in passages, with the task to select the appropriate option during pauses. Before listening to each passage, the participants are given time to read through the question and the options.

II. The second text is an interview with a pantomime artist, divided into five passages with nine MC test items (with four options including one key) associated with it. The procedure is identical with text I: first one listening of the whole, then responses during pauses between text passages.

III. The third part is different from the two previous ones in that it consists of three thematically separate text passages, much like pieces of news presented on the radio. The procedure is, however, the same, with listening of all three passages first, and with responses to four MC items with three options after the second listening.

IV. The fourth and last part differ the most from the previous parts, and also from what the test tasks have traditionally looked like. First of all, the participants listen to each of the six passages only once. They are first allowed to read through the question, and have to select a MC option immediately after listening to the text. The first two passages are short news-flash-like texts, whereas the last four consist of short dialogue situations, where the task is to select a suitable completing speech line among the three proposed.

This amounts to 30 MC items. Each item gives either three points for a correct response one or zero points for a wrong response, adding up to a maximum score of 90 points (of 299 total points of the entire examination of French as a L2).

4.4 A summary of the original test context and the construct of listening comprehension ability

At this point we need to sum up the different details in order to arrive at something like a conceptual as well as an operational construct of the test of listening comprehension. There is no explicit theoretical framework provided by the responsible institution - nor are there any conceptual definitions related directly to the ability being measured. We need to look at the curriculum and to the information provided by the Matriculation Examination Board, and to rely on the experience and knowledge of the constructors in order to define an implicit construct lying behind the test that was devised for the spring 2002 examinations.

First, the test of listening comprehension aims at measuring a general listening ability in the target foreign language, French. The level of the targeted knowledge or ability is defined by a combination of internal and external factors – mainly based on what level is achievable within the limited time devoted to the language at school. In principle, all curricula of all individual upper secondary schools throughout the country are based on the National Core Curriculum. This should be kept in mind, as underlined for example by the Standards for educational and psychological measurement (AERA 1999: 12):

> […] when student mastery of a delivered curriculum is tested for purposes of informing decisions about individual students, such as promotion or graduation, the framework elaborating a content domain is appropriately delimited to what students have had an opportunity to learn from the curriculum as delivered.

Second, there are no clear language functions defined - no TLU situations that should be included in the test contents. The only specific description is “communication in everyday situations”. Third, the thematic fields enumerated in the National Core Curriculum are very broad and varied, and cover almost any imaginable type of text. However, as stated by the constructors, there are practical limitations to the choice of texts – they should be within the interest area and ability level of young people (Cf. AERA 1999: 12, quoted above), and not treat subjects that are too delicate or too topical. Fourth, the possible test formats – related to the operational construct - are defined in the instructions to the Matriculation Examination, but the “best practice” advice on how to create good items seems to be the test constructors’ responsibility. As a consequence, someone not familiar with the educational tradition and the practice of the Matriculation Examination and its implicit principles would perhaps not be able to create good items for the test. It appears to be a question of a subtle feeling for what would be suitable and valid for this particular test purpose. This is a problem, since it has been regarded impossible - because of lack of economic, temporal or practical resources – to conduct any validation studies by piloting the test within the framework of the construction process administered by the Matriculation Board.

In conclusion, what we get is a construct of a general, basic, listening comprehension ability restricted by the target participants' (that is students') plausible level (of vocabulary, grammar and language functions) after a certain amount of language learning in the upper secondary school context, and the participants' areas of interest.

Implicitly and as a common sense argument, I can assume that the spoken language element will be included in the construct. It is indeed stated that the test questions are to be answered on the basis of an audio recording. This means that a test of listening comprehension cannot be based on a written text that is read by the test-taker in order to answer the test questions, with an underlying assumption that a test of comprehension is a test of comprehension, no matter in which form. If that were the case, a test of reading comprehension would include assessment of the listening comprehension ability as well. Buck (2001) mentions this issue, asking why we should give a listening test that is much more complicated to construct and administer, if it does not specifically test the listening comprehension ability. Another invalid assumption would in this case be to say that an assessment of the listening comprehension ability would be covered by a test of speaking, since the ability to understand a language necessarily comes before the ability to produce it. According to this reasoning, then, if a test participant can speak, he or she can surely also understand spoken language. However, with large groups of test-takers as in the case of the test of L2 at the Matriculation Examination, a test of speaking is even more complicated to administer and to score than a test of listening, so this solution is hardly conceivable.

There are an unlimited number of types of spoken texts that could be included in the test based on the concept and implications of communicative language ability. This implies an authentic approach with plausible, realistic situations in which we might find a language learner at a particular level of target language ability. However, as established by the test constructors during the interview, the types of texts selected for the test in reality come from a rather limited range. For practical reasons the spoken texts are most often recreated on the basis of written texts found in magazines or newspapers. There is a shortage of suitable spoken texts that would be on an appropriate level to use in a test. Often, it is simply not possible to find pieces of authentic, spoken text that lend themselves to be used in test items – their length, the voices and noises, or their information structure are simply not convenient. As they would have to be altered very much, creating own unauthentic pieces of spoken text proves more practical. When there is a need to include dialogue situations in the test, they are created by the constructors. According to the constructors, using authentic dialogues would be technically too complicated, and would, moreover, be linguistically too difficult for the test participants.

In order to give a clearer idea of what the targeted level for the listening comprehension test under analysis might be, I will turn to the new National Core Curriculum and the description of the listening comprehension ability. It needs to be pointed out that even if this curriculum was not yet valid in 2002, the level is not expected to be very much different – at least not higher - from the current objective. According to the description of the level A2.2, the learner is to understand simple speech, or follow a conversation on a topic that is of immediate importance for him. Moreover, he or she is to understand the core contents of short, simple discussions and messages (instructions or information) of interest for him or her as well as be aware of the changes in the treated subjects on the TV news. Even for simple messages, the conditions for understanding imply that the speaker uses general spoken language, pronounced carefully at a normal pace. Often repetitions are necessary. This description illustrates the fact that there are clear limitations to what kind of language is expected to be mastered by the learner or test-taker.

On the basis of the background information on listening comprehension processes and strategies, on assessment principles and validation procedures, on the background to the assessment context and the test construction procedure, presented in Part I in this research study, I shall move on to describe the research questions, the data and the methodology applied.

5 THE RESEARCH STUDY

This part (chapters 5 and 6) consists of a description of the research questions, the data and the data collection as well as the use of the methodologies in the study.

5.1 Research questions

As stated in the introduction, the main research questions are the following:

1) What processes are activated and what strategies are employed by the test-takers at seventeen multiple-choice items assessing listening comprehension of French as a foreign language?
2) How does the nature of the individual multiple-choice test items influence the employed processes and strategies?
3) How do the test-takers' listening processes and strategies relate to their success in solving the listening comprehension items?

These are all closely related to the main focus of the study, the information provided by the introspective responses given by the test-takers to a test of listening comprehension of French as a foreign language. In addition, there are some necessary and relevant supporting questions that offer different angles to the validation procedure, namely:

4) What are the contents of the test items?
5) What are the results of the test-takers on the listening comprehension items?
6) What are the characteristics of the items in terms of their difficulty and discrimination?

5.2 Data and collection

Within the framework of this study, several different types of data are used. First of all, the starting point is the test of listening comprehension used in the

Finnish Matriculation Examination in spring 2002 as part of the test of French as a L2 [46]. Added to this listening test, the Matriculation Examination Board has provided some descriptive and statistical information on the original test administration.

On the basis of this original test, I used a pool of listening comprehension items. The procedure for the administration of these items is described in the following.

5.2.1 The test procedure for the current research

In my study, I have used three of the original parts of the test (original parts I, III and IV; see chapter 4.3) in the following manner:

- For the first part (original I), I play the text version with the pauses (35 seconds per item to answer and to read through the next item) at the first listening already, asking the participants to select the suitable option immediately. After the second listening, at which point the participants will also have the possibility to change their choice of an option while leaving visible their previous choice, I ask them to justify their choice briefly. (This procedure is described more closely in the section treating the introspection method.) The numbering follows the original 1-11.
- For the second part (original III), I have changed the item format completely. In my study, the participants listen to each passage twice, and either a) answer short-answer questions or b) write a short summary of the text contents. The time allotted for each of the three items (named A, B and C) is approximately 60 seconds.[47]
- The third part (original IV) follows the pattern in the first part: at the first listening, a choice is made that can be changed during the next listening and the introspective reporting. The numbering follows the original 25-30.[48]

The three parts that are re-used amount to seventeen multiple-choice items plus three open-ended items[49].

After the completion of the test, at the end of the test paper, the participants are welcome to provide comments and reactions, which some of them did, much depending on if they felt they had the time or interest to do it.

I have not included the entire original test, simply because I would not have been able to administer the test with the extra introspective task within the 45 minutes allotted in each of the participating classes. Therefore, I chose to leave out the part of the test that was the most traditional in format, and, in that respect the least interesting. I wanted to analyze the parts of the test that had, according to the item analysis provided by the Board of Matriculation Examina-

[46] The Matriculation Examination, the test construction process as well as the original test procedure is described in part I, chapter 4.3

[47] In the present study, this second part with open-ended tasks is not analysed.

[48] The numbering of the MC items follows the original so that comparison between the original test items and the current can be more easily made.

[49] The first page of the test sheet is found in Appendix I.

tion, worked less consistently well. The time limit of one lesson or 45 minutes relates to the practical circumstances of getting French teachers to lend me their groups and to give me their time within their tight course schedules.

There are similarities and differences between the original test administered in spring 2002 and the test version(s) used during the study:

- The time allotted for the test was the same, approximately 45 minutes, which is the duration of a regular lesson in Finnish schools.
- In the majority of cases, the test took place in a language laboratory. If this was not the case, a portable version of a language lab was used in the classroom, allowing the test takers to use headphones. This is also the situation during the actual test: the test circumstances are naturally made as beneficial as possible for the test takers in order to avoid negative bias for some groups of individuals caused by bad test-taking conditions.
- The instructions in the study were given both in writing and orally in the test takers' mother tongues, whereas the original test procedure uses the target language, French. This is so partly in order to be able to use the identical audiotapes independently of the test takers' mother tongue (Finnish for up to 93% and Swedish for approximately 7 % of the participants in the current study). In all classes where the test was administered, I was present explaining how the test would proceed. Before the beginning of the test, I briefly explained to the participants why I was there, what my research was about ("*I want to find out about the processes taking place in a listening comprehension test situation*"). I gave the test-takers all the instructions for the test, explaining for each of the three parts what they were supposed to do. Additional information given only orally included the explanation why they were to leave their first choice of an option at each MC item even if they changed their choice after the second listening. For the second part, I pointed out that there were some inconsistencies with the recording, just to avoid making them confused while listening.

While this current test constitutes one part of the analysed data itself, it thus also serves as a tool for further data collection. The data obtained on the basis of this revised test consist of the test-takers results on each item, as well as their introspective responses on these items.

The sample of test-takers for the research and the current test comes from schools in the southern and south-western parts of the country, mainly from urban but also from rural schools. This was in order to represent, as closely as possible, the actual target population for the original test context. The 218 test-takers thus consist of upper secondary-school pupils in 22 different schools in the following towns or municipalities in Finland: Helsinki, Jyväskylä, Naantali, Nousiainen, Paimio, Parainen, Raisio, Tampere and Turku. The municipalities, the schools as well as the classes to whom the test was administered were of very differing sizes. The pupils are assigned codes, so neither the individual pupils nor the schools can be identified in the study.

The test-takers were selected based on the willingness or the possibility of the individual teachers, contacted personally by e-mail, phone calls or mail by the researcher, to let her or his pupils take part in the test. Approximately 90 %

of the teachers who were contacted agreed to receive the researcher in their school to administer the test during French lessons. A few teachers refused to cooperate due to lack of time or interest.

The data collection took place in the years 2004 and 2005, in situations where the pupils were to take part in the Matriculation Examination within the months that followed. Participation was on a voluntary basis for the pupils, with the assumption that the test would be a useful practice. Some of the teachers wanted to obtain their pupils' results from the test, and were also provided with them.

The numbers and proportions for both the original and the current test administration are as follows:

TABLE 15 Number and proportion of test-takers in the original and the current test

	Finnish-speaking schools	**Swedish-speaking schools**	**Total number**
Number & Proportion of test-takers in original test 2002	2983 91,5 %	279 8,5 %	3262
Number & Proportion of participants in current test	203 93 % (20 schools)	15 7 % (2 schools)	218 (= 6,5 % of the number of participants in the original national test)

Another type of data was collected by means of an interview with two test constructors. The information obtained serves as background information to the testing context due to the lack of any explicit detailed test specifications for the Matriculation Examination, and are reported on in chapter 4.2)

5.2.2 Remarks on the test procedure and the study

Some general remarks can be made about the test procedure and circumstances from the point of view of the participants in the study, and about the limitations to the current research procedure as a study validating a test item (use), since an actual complete test of listening comprehension would not look exactly like the researched test.

Naturally, the current test situation is different from the original high-stakes situation, where the score obtained on the listening test is added to the scores on the written part of the examination, resulting in a total score for the participant's French language ability. Motivation is probably higher among the participants in the original test situation, as their examination results are at stake, but so is probably the experienced anxiety. The participants are probably less motivated to make an effort at their performance in the study situation, even though many do their best to profit from the test as a way of practicing for the test that they will take in a near future. As can be judged by most of the teachers' willingness to participate in the study, and by their comments in the test situation, the pool of items in the study bears enough similarities with the original test items to be able to function as a useful exercise and to work as a diagnostic tool both for the students and their teacher. The texts and the items constitute parts of a real test with the likelihood of being similar to test items in the future real test situation. Therefore, the motivation factor is not as low as it would be for a test with no real-life context or consequences for the participants.

At the same time, I have deliberately wanted to create a relaxed atmosphere for the test situation, in order to make the test takers react as spontaneously and abundantly as possible to the introspective task. I am aware of the fact that this creates a contradictory situation, where I want the participants to perform as closely as possible to their real listening test-taking skills, but at the same time to be comfortable in order to react freely to the task. The best way to minimize invalid effects is to have as many participants as possible, in order to maximally level out the effects of the test situation. Some participants will be more affected than others by these new circumstances – in either a positive or a negative way.

Most of the participants have practiced for the actual test ahead of them before the current test situation: the amount of time spent on this particular test format in class, using tests from earlier years, varies from one teacher to another. All participants are, however, familiar with the specific test format used within the framework of the institutionalized Matriculation Examination, often both from their target language French, and from other language studies (two other compulsory[50] languages: usually English and the second national language, Swedish or Finnish). Therefore, their test-taking habits are very much marked by these earlier experiences. Some participants are more sensitive to changes in the test taking procedure and formats than others, and the most sensitive will naturally be affected – most often probably negatively – by a change in the test situation. This is an interesting issue in itself, concerning the validity and reliability of tests and the effect of format, preparing for a test, washback and authenticity. Real language ability should not be situation dependent, and especially not test task-dependent. In practical situations it often is, however, depending on a learner's experiences of situations and tasks in and outside of class, on test-taker personality and similar factors. This also relates to the effect and the authenticity of the teaching. Ideally, the test task should be so transparent and clear, that the targeted ability can be reliably extracted from the test-taker's results on this test task.

The fact that the test procedure is changed from the usual procedure does, however, influence the test-taking process. In the first part of the test (items 1-11), the first selection of options has to be done already after the first listening

[50] To pass the examination in these two languages was compulsory at the time of the original test and the study. Today the only compulsory subject is a test of Finnish or Swedish as a native language, added to three other subjects selected among the available subjects (the second national language (Swedish or Finnish), a foreign language, mathematics or science). It is, however, possible to add as many tests as wished by the test-taker. (L 766/2004, 18 §)

and that is a complicating circumstance. The participants have much less a chance of getting the context or the theme of the text before they have to start drawing inferences on the basis of a short text passage. This mainly affects the students who are normally able to use the context in a holistic way to get clues as to the meaning of details. This is a useful (meta)cognitive capacity and a part of the construct of listening ability where understanding of larger chunks of speech is aimed at. Another part of this construct also covers the understanding of separate text fragments. In a real-life listening situation, the listener usually has a clue as to what the larger context is – be it a radio emission on tribal habits in Borneo, or an announcement of the arrival of a train at a railway station. Nevertheless, as judged by the test results, many students seem to know how to profit from the fact that the text passages are heard a second time – they are able to reconstruct the context by adding up the separate pieces of the spoken text. With the current test procedure, if the students think it is necessary, they can reconsider their interpretation of the text and change their selection of options accordingly. An advantage with this procedure is that the participants have the possibility to consider the items – stems and options - more carefully, since they read through them twice. This may, in fact, be a disadvantage for the test validity as the test-takers' focus moves to the written part of the task thus possibly causing construct-irrelevant variance. This is related to the problem of the MC test format: the options should be transparent enough not to cause comprehension problems for the participants or interfere with the text interpretation process. It is the degree of comprehension of the spoken text that should cause variation in test scores, not the comprehension of the written question.

The second part of the test constructed for the study uses the original spoken texts, but treated in a slightly different way. The original procedure consisted of listening to three thematically separate passages all in one go, and then divided into passages, at the stage of responding to MC items. I decided to try out a different test format in order to investigate the effects it has on comprehension. Therefore, I used the short-answer format, where I let the participants listen to each passage twice before answering a question related to the text or write a short summary. I tried to keep the original MC question wherever possible without providing the options, changing the language though, so that the question and the answer given by the test takers were in their mother tongue, in order not to cause any cognitive load related to writing in L2. Unfortunately, the quality of the audio tape recording was not optimal. The pauses between the two chances to listen were uneven, and some disturbing noises could be heard. This should not have interfered with the listening of the text itself, but may have annoyed or confused some students. Another source of confusion was the fact that there were no taped instructions for this part of the test, as there was for the other two. Since I had changed the procedure completely for this part, I could not use the original instructions on the tape. I gave them explicitly at the beginning of the testing situation, but the participants' ability to take in all that amount of information appeared to vary. Some were confused at this part of the test, which may have interfered negatively with their entire test-taking process.

Moreover, the test task format with open-ended questions was, contrary to the 'institutionalized' MC format, unfamiliar to most of the participants, even though some might have come across it as it is used in the Matriculation Examination in listening or reading comprehension tests in other languages. This might also affect the participants' performance (in unpredictable ways).[51]

The third part of the test consists of the six last items (25-30) of the original test. As opposed to the original test procedure, which consisted of listening to each separate passage only once and then responding to one MC item per passage, the procedure of the first part of the study test was repeated: the test-takers got a second listening and a chance to change their option selection at the introspective stage of the task.

It is evident that at all stages of analysis of the test constructed for the study, the differences in the test contents compared to the original test have to be kept in mind, especially when making inferences about the relative validity of the use of the original test items. In order to validate the current study, I shall assume that the differences are not of a kind that would make the two tests and the circumstances around the test administrations completely incomparable.

By a close comparison of the two administrations of the selected seventeen test items, the proportions of the choices of an option for each individual item, not only for the key but also for the distractors, are considered. The null hypothesis implies that there is no significant difference between the results in the original test and the current test items, as far as the proportion of choices of options for each item is concerned. By applying a chi square test (by means of the SPSS 17.0 software) I was able to establish that for 12 items (items 3, 4, 5, 7, 8, 10, 25, 26, 27, 28, 29 and 30) the null-hypothesis is supported (Exact sig. >.05) whereas for 5 items (items 1, 2, 6, 9 and 11) the null hypothesis is rejected (Exact sig. <.05) item 9 being very close to the set limit (Exact sig. .049) (See Table 16).

[51] This second part of the test is not treated, analysed and discussed, within the framework of this study, but the data obtained lends itself to a type of study where the plausibility of the MC options can be investigated.

TABLE 16 Chi-square analysis of 17 items: results from original and current test administration

Item	Chi-square	Exact Sig.	Null-hypothesis supported
1	6.903	.032	
2	24.089	.000	
3	.379	.837	X
4	2.879	.235	X
5	5.626	.060	X
6	9.546	.008	
7	2.431	.297	X
8	1.410	.491	X
9	6.025	.049	
10	.963	.628	X
11	6.405	.040	
25	3.652	.164	X
26	1.371	.509	X
27	3.603	.168	X
28	1.113	.565	X
29	1.327	.528	X
30	5.209	.077	X

In Table 17 (in Appendix 2) are presented the proportion of option choices for each item for the two test administrations (original and current). Most of the seventeen MC items have been more difficult for the current test-takers than for the original test-takers. If we consider items 1, 2 and 6, for which the null hypothesis in the chi^2 analysis was the most clearly rejected, items 1 and 2 have been much more difficult (with a difference for the key of 8 and 15 percentage points respectively) in the current administration compared with the original administration. It can be speculated that the fact that both items 1 and 2 are the two first items in the new test with new tasks for the test-takers - the introspective reporting – makes them likely to function differently and be more difficult for the test-takers in the research context. For item 6, the proportion of test-takers who have selected the key do not differ very much between the two administrations (2 percentage points), the main difference being in the distribution of choices between the two distractors (with a difference of 9 and 7 percentage points).

For item 9 there are differences mainly in the figures for the key (8 percentage points), slightly less for the two distractors (3 and 6 percentage points). Item 11 have nearly equal proportions of selections for the key, while the proportions for the two distractors differ only with three and four percentage points.

Based on the figures reported above we can conclude that 70% of the items worked in a clearly similar manner, and 88 % of the items in an almost similar manner in the two testing events. This will justify a comparison between the two test administrations. However, the different test circumstances in the two test situations (the original high-stakes examination context and the current research study), described above, combined with the smaller number of test-takers for the research study have affected the results of the test as well as the chi^2 analysis, to some degree. Some explanations for the quantitative differences described above are expected to be found in the qualitative analysis of the contents of the items (in chapter 7) as well as in the introspective responses (reported in chapter 9). Further quantitative analyses of the results for the current administration of the test items are found in chapter 8.

6 METHODS OF ANALYSING THE CURRENT TEST ITEMS

As has been pointed out in the introductory chapters, there are mainly three sources of evidence of the quality of the listening comprehension items in the study. To this adds the information on the test construction process (presented in chapter 4.2) that provides both a background to the present construct and can be taken as a source of validity evidence.

First, the item contents needs to be analysed in relation to the construct. Second, a Rasch analysis as well as a distractor analysis is undertaken in order to get a quantitative justification for the content and internal test analysis. The item measure and discrimination as well as the person measure are established. However, the principal source within this study consists of the test takers' response patterns revealing internal processes by means of the method of introspection[52].

6.1 Ways of analyzing the test contents

As far as the analysis of the test contents is concerned, in this study, the aspects to take into account mainly follow the principles of AERA (1999), Weir (2005) and Buck (2001) in their description of content or context validity (See chapter 2.7). Thus I will take into account features like the themes of the text, the tasks, the item format and test administration procedures.

Content validation typically involves experts making judgments in some systematic way (Alderson et al. 1995: 173). It should be pointed out here that expert judgments have already been made on the contents at the stage of creating the test items and constructing a test to be used in the actual test situation in the Matriculation Examination in 2002. This justifies the fact that for the current

[52] The theoretical framework for this methodology is described under chapter 3.

study I represent the experts alone, sharing the roles of a test item writer, language teacher and researcher.

It is fruitful in this context to consider elements in the spoken text passages and the questions related to these passages (where the notion of *necessary information* – the textual information needed to be able to answer an item (Buck 2001: 129). – becomes essential). What kind of vocabulary and structures as well as discourse features do the text passages contain? What language functions are assessed? Are these central for the target audience and the targeted level of the skill – how do they relate to the construct? How do these affect the difficulty of the comprehension task?

Some studies consider quantitative features in the test contents (in the spoken text and in the interaction of the text and the questions) and correlate these with test results: text length, placement of necessary information, number of propositions and speech rate (See Rupp, A.A, P. Garcia & J. Jamieson 2001: 195-6). While these characteristics do contribute to the general contents of the test, giving information on its potential level of difficulty, it is difficult to isolate and separate them to make judgments of their influence on test results. Moreover, since in this study these details are not controlled, as we part from an existing product, there are no points of reference or different items for the purpose of comparing variables. Therefore, in this study the features of the test contents in different text passages are treated more as a whole, and speculated upon from an external expert point of view. However, some essential detailed traits will be focused on.

6.2 Statistical information on the test

By means of the use of Rasch analysis (*WINSTEPS* 3.69.1) and other analyses obtained by means of the software SPSS 17.0, some useful statistical information on the individual items is obtained. This information is compared with the information collected by the Matriculation Examination Board on the original test administration: the proportion of correct and incorrect responses as well as the attraction of each of the options (see chapter 2.6 on the analysis of MC items and Table 17 in Appendix 2).

The descriptive statistics of interest include the mean, the median and the standard deviation of the test-takers' results on the test as well as on each item. The Rasch analysis provides information on how the test items have worked in relation to each other and for this specific group of test-takers. I will treat this information in relation to quantitative information on the introspective responses, completed with graphic item and distractor analysis.

The quantitative information serves as background information to the more profound qualitative analysis that serves as explaining information. The introspective analysis indicates where there might be possible flaws or points of interest and possible needs of revision. The ultimate goal is to profit from all the available resources in order to see how the validity of the item score applications is potentially affected by different test characteristics and circumstances.

6.3 The use of introspective analysis in the current study

The ways of employing the introspection method in the present study are described in the following. Based on the categorisations and definitions by Faerch & Kasper (1987), Green (1998) and Cohen (1998a) described above (chapter 3.1.2), the type of verbal protocol used in the present study can be summarised as in the table below (Table 18).

TABLE 18 Characteristics of the verbal protocol used in the current study

Dimension	Current study
Type of protocol	Self-observation or self-revelation related to a specific situation and task
Type of object	Cognitive/affective information
Temporal aspect	Immediate retrospection related to each item
Role of experimenter	Non-mediated protocol, examiner passive, only providing instructions in the beginning
Data collection instrument	Open-ended; Written output; Time and writing space limited
Training of participants	No training

The innovation in the application of the introspective method in the current study is the procedure where the test-takers briefly write down their thoughts. This happens in an indirect way as the test-takers are asked to justify their selection of an option at each item. The advantage of this procedure is that larger numbers of test-takers can take part, allowing for a possibility to correlate quantitative data with the qualitative information on each item obtained through the introspective method. Cohen (1998a: 29) talks about a major benefit of large-scale surveys being the potential for generating and testing hypotheses because of the large number of respondents. The disadvantage for the type that the current study represents is that the information obtained from and of each individual is scarcer, and perhaps less profound, since space and time are only provided for very short responses.

The participants in this particular study do the test without any training, so indicating the reasons for their choices is an entirely new activity for them. As might be expected then, while some participants find it natural, easy, useful and interesting, others do not. Moreover, it has to be kept in mind that there are activities at three levels present at the introspective method:

1) The unconscious thinking processes and the more conscious test-taking or comprehension strategies applied when listening and solving the task.
2) The giving of a response, i.e. the selection of an option.
3) The introspective responses.

For the purpose of the analysis, my assumption is that levels 2) and 3) reveal something about the activities at level 1).

While it is clear that the supplementary cognitive activity of the task, i.e. stating the reasons for selecting a particular option, does demand more effort than a regular test situation, putting on cognitive load to the situation, the task is "de-dramatized" by 1) not asking the test-takers to explain what goes on in their mind – which would be far too demanding and time- as well as space-consuming – but on a more concrete level, simply asking them to state why they have selected a particular option; 2) not giving them very much time nor space, which should be indicating that it is not very much that is asked of them and 3) this added task not being compulsory, i.e. allowing for empty boxes in case the participants feel it is too demanding to simultaneously concentrate on both the item solving task and the introspection.

Following the recommendations of Allan (1992), Cohen (1998a) and Goh (2000) I have also chosen not to ask participants to specifically name the strategies they have made use of – in my study it would have demanded far too much practice and explanations about what the particular strategies actually imply in order to be able to reach any important conclusions. The exception is the indication of guesses. I have thus taken the risk of letting participants react rather freely upon the stimulus and task, thus also permitting new, unpredictable reactions to arise.

The outcome of the use of this particular short written introspective method gives at hand the fact that even though the participants have experienced that the task of writing justifications for their selection of options has been demanding (as judged by many comments given in a space provided for this purpose at the end of the test paper), most participants have been able to write justifying responses – just one or two words, perhaps, but in some cases also very elaborate responses.

It may be the case that for those test-takers whose success partly depends on the fact that the test procedure follows a certain, familiar pattern, the added task influences the test situation the most. This familiarity with the test procedure also includes the time given to respond – with an extra task, the time gets scarcer, and for some learners, this pressure affects their performance negatively.

Coding and categorizing the introspective responses

A coding scheme should be understandable and usable for individuals other than the developer of the scheme (Green 1998: 97). The primary validity issue in this context is the extent to which the categories and codes that are used to analyse the information reflect the reported processes. The reliability concerns the consistency of the coding procedures and categories. Bachman (2004: 279; cf. Green 1998: 94) recommends that a second independent coder code at least a proportion of the segments that are to be coded by the principle coder, as assessing coder reliability is an important phase in protocol analysis. The interrater consistency for the present study is verified with two other independent coders, who have coded all the responses given by the 218 participants for one of the items (for results see chapter 9.2).

The fact that there are no lists and thus no limitations to the kind of introspective responses that the participants will provide might naturally cause problems for the researcher at the stage of gathering the information obtained from the responses, since they are rather varied (Cf. Cohen 1998a: 28). The categories in the current study (presented in chapter 9.1) are based on both pilot studies with a similar test format (Anckar 2003) as well as on partial results in an analysis with a limited number of test participants' responses collected. Some of these categories have been simple to label from the beginning, whereas other types have been renamed and re-categorized several times. As the method applied in this study is, to some extent, completely new, I have not been able to use earlier lists of response types directly as basis for my categorization. However, the bases for the applied categorization in the present study are related to the assumptions that the processes and strategies activated or deliberately used at a listening comprehension test-taking situation are dependent not only on the spoken input, but also on the test format and other features in the test setting. This is expected to be reflected also in the introspective responses that the test-takers provide.

The objectives for analyzing, describing and categorizing the introspective responses are two-sided. First, there is the aim of describing the quality of individual items. The question to be answered is what the introspective responses given at a particular item can tell about this item. An example is the mention by a majority of the test participants having faced unknown vocabulary at use in the options for a particular item. In this hypothetical case the results in the tests for many participants are thus influenced by unknown vocabulary in the written options, which may not belong to the listening comprehension construct proper. This is naturally valuable information, as it reveals possible traits in the test that adds to skewed results and, consequently, construct-irrelevant variance.

The second aim of this categorization is to compare the introspective responses of participants with different levels of success in the task of solving the listening comprehension items. From the viewpoint of test construction and item validation, it is essential to know the patterns of the processes of both weaker and stronger test participants, as the items should give truthful and reliable results for all participants. Therefore, it does matter if the listening comprehension items test different things depending on the ability of the participant. The most covering picture of the test process is obtained by combining the information from these two sources: the items and the participants.

ANALYSIS OF THE LISTENING COMPREHENSION TEST ITEMS: RESULTS

In the third part of the thesis, I will report and discuss the results of the analysis of the item contents, the test-takers' results at each item as well as the introspective responses.

7 A PRIORI ANALYSIS OF THE CURRENT ITEMS

Before moving on to the analysis of the test-takers' results in the current test and the main analysis in this study, based on the introspective responses, there are features related to the test contents that need to be described and analysed. This analysis parts from the knowledge of the fact that the items have already been treated and worked on by the French division and the language board of the Matriculation Examinations. The content analysis is thus based on the cognitive theories of the listening processes and principles for creating good test items as well as the context for the present items and the original test of listening comprehension.

The contents can be considered as a whole: what type of text, theme and tasks in general are at stake in the test? Moreover, separate characteristics related to individual items need to be considered, in order to be able to relate the results and introspective responses to the big picture of the assessment context and the present construct. There are some items where specific phonological details seem to be at stake. Textual questions related to the content validity are for example: Is the theme of the text convenient, i.e. of interest, sufficiently familiar, but not biased, for the intended test-takers? What about the factors of lexical difficulty - does the text include technical jargon, abstract notions, idiomatic expressions or other potentially problematic vocabulary? Are the syntactic patterns on a suitable level of difficulty? Are the language functions clear enough and within the potential ability of the test-takers? In the following, the contents of the text are described in relation to these questions: I will now proceed with a more detailed analysis of the test contents on the basis of the text, the test tasks (items) and their interplay.

7.1 Items 1-11

As this first part of the test, covering items 1-11, consists of a whole, but with five passages with two or three related items, I will proceed in the following manner: first I will describe the text contents of part one as a whole, then go to the details of the two or three items that relate to each passage, at that point focusing on the item format and tasks.

Some of the current test-takers' teachers criticize the theme of this text. We can assume that the teachers' judgment, even though it is based on subjective opinions, is fairly valid, since they work with the students continually. Thus they have probably, through their experience after years of teaching, got an idea about what types of text themes may be of interest to adolescents. Since most of the students live at home and go to school, issues related to buying apartments, redecorating and moving may not be of top interest in their lives at the moment. However, one could argue that this is a text that gives an idea about the way people live in France, which should be interesting information for anyone studying the French language including the culture associated with the target language. The text content does not seem to be biased towards any particular group of test-takers – on the limit we could say that urban adolescents might have a clearer picture of the way small shops and workshops used to be placed in towns. However, since the exteriors of buildings in Finnish towns in general differ from those seen in France, this is not a very good argument against the suitability of the theme of the text.

The vocabulary varies from more abstract, but not particularly technical or specific jargon, using words and expressions like *phenomène* 'phenomenon', *la vie communautaire* 'community life', *hésiter* 'hesitate', *la motivation économique* 'economic motivations', *un acte volontaire* 'a voluntary action' *un esprit d'ouverture et de rencontres* 'spirit of openness and meetings', *l'inconvénient* 'inconveniency', *originaires du Midi* 'come from the south of France', *au bout de mes forces, crevée* 'completely exhausted', *surveiller* 'keep an eye on', *à l'aise* 'at ease', *la solitude* 'the loneliness' to rather concrete lexical elements : *boutiques* 'shops', *appartments, campagne* 'countryside', *table, trottoir* 'pavement', *poubelle* 'dustbin', *perdre plusieurs kilos* 'lose several kilos', *clés* 'keys', *complet* 'full', *enfance superbe* 'superb childhood'. As we have seen in the description of issues related to the vocabulary within the listening framework, and as for example Biber (1995), Weir (2005: 74) and Rupp et al. (2001) point out, abstract information is cognitively as well as linguistically potentially more complex to process, and that has to be kept in mind when considering potential difficulties at the items. This aspect is also taken into consideration for instance in the *CEFR* ability level scales.

What is essential to consider is the interaction between the text and the task. The difficulty of the text contents is dependent on what the items look like, that is, what questions are asked. A useful concept in this respect is "necessary information" - the information in the text that a test-taker must understand in order to be sure the task has been done correctly (Buck 2001: 129). According to this, not all units of vocabulary in a text are necessary to understand, or even to catch, when listening. A lot of information is redundant; it might be repeated or paraphrased, or simply not essential with respect to the asked questions. Bearing this in mind, I will come back to details in vocabulary when treating passages associated with each item.

The information structure of a text is very important from the point of view of comprehension (see chapter 1.5.1). An important issue to discuss here is the fact that the text presented in this test in the spoken mode was originally written. It is actually a complicated case, since the text is authentic, originally probably based on oral interviews, but it is written down to be used as an article in a magazine. The original text has been modified to suit the purposes of the test and the language level of the test-takers, and then recorded by actors reading out the text (See description of the test construction procedure, chapter 4.2). Therefore its information structure is partly typical of a written text, partly preserving some characteristics of the original oral interviews. The text can thus probably be placed somewhere in the middle of the oral-literate scale (Tannen 1982) exhibiting moderate degrees of spokenness and writtenness (cf. Flowerdew & Miller 2005). We get, as a result, an argumentative text, where a main claim ("Today people live in shops in Paris") is followed by supporting facts and reasons leading to the claim ("Small businesses close down", "People are not afraid of the ground floor" "Ordinary apartments have become too expensive", "The housing agencies don't hesitate to sell old shops"), and with added "real-life" evidence and arguments for the truthfulness of this claim. There are several opinions presented by different people, advantages and disadvantages to the presented phenomenon.

From the point of view of test construction, this type of text seems to lend itself well for creating items, since it is possible to divide the text at natural boundaries, at the changes of speakers. The text might even have been possible to use in its original oral presentation mode, since the structure seems rather clear in itself, and the text would have presented the chance to include some regional variants in the speech lines of the different people speaking in the text. However, due to the limitation of the test length and complicated constructions in the original text, this is impossible in this test context for test-takers on this target proficiency level.

The first passage is an introduction to the phenomenon, and the rest represents the speech turns of different cases, representing nearly independent texts. The nature of the text clearly lends itself to the format where the text is first presented in its entirety, in order for the test-takers to get the overall picture and the main subject or claim. Then, at the second listening, it is motivated to go to understanding details. However, even if this was the case at the original test administration, because of reasons explained above, this was not the procedure for the study.

An aspect to take into account, related to a larger practical and philosophical issue of the contents of teaching, is the authenticity of the text in respect to being an authentic piece of spoken language. There are two issues at stake: first the contents of the curriculum, and second, the test specifications. If we look at the recommendations given in the National Core Curriculum serving as a theoretical framework for the creation of the listening construct for this particular limited purpose, we can conclude that what is missing from the type of semi-spoken language in the test, is the "everyday language spoken at normal pace", the expressions and structures typical of that language, and the strategies needed to be able to handle this kind of language. Instead the language of the test represents more the typical everyday – written - foreign language encountered in class. This is fair, since if the authentic or genuine everyday language is not frequently presented to the students, neither can it be tested (Cf. AERA 1999: 12). Therefore, the test follows the implicit, hidden, actual curriculum present in class. One can only speculate on the effect on the teaching in class of letting the test actually consist of authentic spoken texts. After the first chock and protesting and possibly rather bad results for the first group of test-takers, there would probably be a positive (from the point of view of the default listening construct) washback –effect, consisting of a higher frequency of authentic pieces of spoken language presented and worked on in class, which would lead to learners experiencing more confidence in facing the foreign language in its spoken mode outside class. This would follow the principle ideas of the communicative, task-based and learner-strategy oriented approaches to teaching a L2 (see chapter 1.2). Taking account of the characteristics of the type of language and text content of the test, we are obviously situated away from the default listening construct proposed by Buck (2001), within which he points out the necessity of including as many as possible of the elements specific for listening that cannot be tested otherwise.

At this point it is purposeful to move to look at the items in detail: we are interested in the text, the task, and the interaction between these two. What are the demands on the test-taker, what potential difficulties do the items present, what construct-irrelevant factors can there be?

Identifying text variables that consistently cause difficulty is a complex task (Alderson 2000: 70-1). Clearly at some level the syntax and lexis of texts will contribute to text and thus item difficulty, but the interaction among syntactic, lexical, discourse and topic variables is such that no one variable can be shown to be paramount.

Generally, we can say about the potential difficulties that a test-taker may experience, that they represent an interplay of several different factors ranging from phonemic discrimination and segmenting problems, to problems with larger units related to everything between vocabulary and discourse structure (according to Dickinson 1987 and Brown 1995 quoted above, chapters 1.3 and 1.5.2). It is difficult to pinpoint any specific details at each listening text passage that may cause trouble for any individual. The problems related to the French language in its spoken mode that the learners and test-takers face partly have to do with the way the discourse consists of segments with boundaries at other than word boundaries. The fact that letter combinations are pronounced in certain very different ways from what is the case with the mother tongue of the test-takers, the frequency of elisions, liaisons and other *sandhi* –phenomena adds to potential problems (see chapter 1.5.1). However, when treating these text passages and their comprehension, I will focus more on the information content, the argument structure, the discourse coherence and other larger units, thus not entering into possible difficulties on the phonological level. The task for the test-takers at this level is, after all, to understand larger units than separate words or discriminate between sounds. These abilities are included in the higher-level skills at stake in the current test.

The MC item format is traditionally frequently used for large-scale assessments above all because of its ease and objectivity in scoring. However, many critical voices have been raised against the MC item (see chapter 2.4.1). The regular use of more than one test method for testing any ability is recommended (for example by Alderson et al. 1995 and Buck 2001). Moreover, for series of institutionalised tests that are given from year to year, varying test methods year by year will reduce the predictability of the test's format, and possibly the learning of test-taking strategies for particular test methods (Alderson et al. 1995: 46).

Another characteristic related to the test format is in this first part of the test, where two items are related to one passage, and both are to be answered within the same limited pause. In such a testlet, there will obviously be effects from one item to the other. There is a lot of spoken information to be handled in short term memory while searching for information in long term memory for the interpretation of the text, as well as while processing the questions and options related to the two items. These circumstances have to be taken into account when considering the item statistics and the score patterns. Here is an example: if most of the processing capacity and thus the allotted time is taken up by item 1, there will be less capacity left over for item 2. This might consequently – additionally or mainly - affect the scores at item 2. In other words, a difficulty in one item may be reflected in the scores of the following item. The defence for this test format can be the assumption that if the test-taker understands the passage as a whole, there will not be any difficulties with responding to the questions that are aimed at testing the understanding of the key contents in that passage. This is naturally true, but only if the questions and the options are clear and transparent $^{\text{enough}}$ not to increase the test-takers processing load. Otherwise we might arrive at a situation where unclear, problematic written options influence the test-taking process for not only one but two items, leading to construct-irrelevant variance, obviously with unfair and negative consequences for the test-taker on one hand, and unreliable item results on the other.

In the following, traits in the text-item interaction are considered separately for each testlet and item. The approach is one of the researcher's own expert judgement, based on knowledge about the learning background of the test-takers.

Items 1 and 2: content analysis[53]

This passage introduces the theme, presenting the phenomenon treated in the text to come. As the test-takers are allowed and, indeed, supposed, to read through the written questions and options related to this passage, they obtain by doing this a certain idea about what the passage will treat. This is important

[53] The item questions, options and the spoken text passage with translations are found in Appendix I

from a cognitive point of view, since the question and options will most likely shape the listening process (cf. the model of information processing by Jamieson 2000, Table 7 and the issues related to item preview, in Yanagawa & Green 2008, discussed in chapter 2.6.1 above). Words and expressions in the options that are most familiar to the test-takers will be most easily retained in short-term memory, in order to prepare for the processing of the text to be listened to. Such familiar words presented in the options are probably *boutiques* 'shops', *prix* 'price', *argent* 'money', *trop cher* 'too expensive'… Contrary to this, *logements* 'flats', *rez-de-chaussée* 'ground floor', *lien sociaux* 'social relationships' are expressions that might cause comprehension problems for weaker test-takers, which is very problematic, since they are key concepts in this passage. The question from the point of view of validity is whether these are words that should be supposed to be familiar to all test-takers at this level, i.e. whether they belong to background knowledge shared by everyone and thus not influencing the test outcome for any individual? If we get differences in the results of the test-takers because of the fact that certain lexical elements in the written options are not understood, there seems to be a case of construct-irrelevant variance (construct- irrelevant difficulty in Messick 1989: 34). This problem has to be verified by analysing the test-takers' introspective responses.

In question 1, the stem helps the test-takers getting oriented in the text, as the term *phénomène* 'phenomenon' is given in both the stem and the text. The task for the test-takers will be to verify which of the contents of the options corresponds to the contents of the text. Option 1a will probably attract many choices, since one of the key words, *boutiques*, a highly frequent word, is present in the text. Moreover, for those who can understand the paraphrase *"Les petits commerces ferment"* 'The small businesses close down', as it is the first reason stated in the text, this option will be equally attractive. 1b is a summary of the main theme of the text. There is no direct explicit correspondence in the text, but the test-taker has to draw on many statements that together will sum up to this conclusion. 1c is a combination of two elements in the text, a characteristic "*prix élevé*" 'high price' found together with another word-match in the text: *rez-de-chaussée.* The nearly synonym expression of the characteristic in the text *trop cher* 'too expensive' is, nevertheless, characterizing a different object in the text.

In question 2, the verb *explique* 'explain' is matched in the text, so locating the needed information is facilitated. The problem lies in knowing which of the several pieces of explanations in the text can find a corresponding option. To select 2a, the statement *"la motivation économique est évidente"* 'the economic motivation is evident' seems to be dominating. In order to reject it, the rest of the phrase *"mais…"* 'but' has to be taken into account, falsifying the adverbial expression "*uniquement*" 'only' in the option. In 2b the word-match *trop cher* together with reasoning based on real-life experience may lead to test-takers choosing this distractor. A facilitating detail for the selection of 2c is the fact that *sociaux* is matched with *social* in the text, presented between two commas, and therefore pronounced very clearly. *Liens sociaux* 'social relations' in the option is paraphrased *"On cherche la vie communautaire…esprit d'ouverture et de rencontres"* 'People look for community life… spirit of openness and meetings.'

To sum up, the first two items may be characterized as testing the ability of looking for information on the main theme of the text, rejecting secondary themes and reasons, demanding text-based inferences, and adding of separate pieces of information. The abstract notions and factual point of view adds to the difficulty. The main type of distractor is one with matching units of vocabulary, but with false combinations of these units compared to the text contents.

Items 3 and 4: content analysis

This passage contains the first testimony of a person experiencing the phenomenon introduced in the previous passage. The speaker describes the way he works and his surroundings. There is a number of descriptive vocabulary and vocabulary expressing attitudes towards things that need to be understood, and which serve as a basis for necessary inferencing: *genial* 'fantastic'*, l'inconvénient* 'inconvenience', *tranquille* 'calm', *désagréable* 'unpleasant', *touristique* 'touristy', *bizarre, m'ennuie* 'annoys me'*, à l'aise* 'at ease'… The questions are related to the speaker's attitude towards people.

Item 3 asks how the speaker, Joël, describes his life. The question is very vague and does not help the listener localize the necessary information, because it could be anywhere in the passage, since everything can be interpreted to be a description of his life. As a consequence of the fact that the options do not rule-out each other, the test-taker has to treat all the three as true-false statements. In other words, logically, all three statements could be true at the same time! This leads to the case where practically everything in the passage is "necessary information". Option 3a gives an opposite to the inferred text contents. There is, however, a near-match to *"touristique"* in the text, so this might lead to weak listeners choosing this option. 3b is interesting in that *chômeurs du quartier* 'unemployed in the neighbourhood' in the option is connected to the near paraphase *"de gens qui ne travaille pas dans le quartier"* 'people who do not work in the neighbourhood' in the text. The option is clearly falsified, however, by the use of the verb *s'occuper* 'take care of' in the option with no semantic relation to the text contents. The selection of 3c, the key, demands an overall understanding of the situation presented in the text. No direct explicit word-level clues are given.

The stem in item 4 asks the question how people behave with the speaker. Information from different parts of the passage need to be put together and contrasted with the options. There are clear paraphrases in the text pointing to the truthfulness of option 4a, the key: the test-taker has to understand what the speaker's attitude towards people around him is. Option 4b plays with the verb *photographier / prendre un photo* 'take a photo' and the idiomatic expression *prendre pour* 'take for', which has to be understood in order to be able to discard this distractor. Option 4c includes a near-word match with *touriste - touristique* that might attract weaker students. Stronger test-takers may reflect on whether *les gens me photographient comme une personnalité bizarre* 'people take photos of me like of a bizarre personality' corresponds to this option, but will, if having enough processing capacity, notice the mismatch between these contents.

The two items 3 and 4 contain an assessment of the comprehension of the speaker's attitudes towards a situation or certain circumstances and of the ability to match the overall meaning of the text contents to several options.

Items 5 & 6: content analysis

This text passage contains a description of a person's experiences of buying an apartment and the problems related to it. In fact, her tone of voice reveals a frustration, and a negative attitude towards the situation. It is only at the last sentence, that the speaker sounds happier.

The question in item 5 is very vague, and can apply to everything the speaker says. There is no help given as to the localization of the key contents in the text. However, the options indicate that the question concerns the speaker's feelings towards the situation. The options could logically all be true at the same time. 5a contains a negation, a modifier usually said to consume more capacity to process than an affirmative statement (see Freedle and Kostin 1999 and chapters 1.5.1-2). The task for the test-taker is to locate the subject "*les bruits* "the noises', and find out the speaker's attitude towards it. There is a near-match in the text with *"trop bruyant"* 'too noisy' which implies that the noises do disturb her. Therefore 5a should be interpreted as a distractor. At 5b potential problems may arise with the verbal expression of time "*vient de*" (literally 'comes from'), indicating something that has happened in the near past. *Problèmes de santé* 'health problems' has to be associated with two stated facts in the text: *J'ai fini par me sentir au bout de mes forces, crevée! J'ai perdu plusieurs kilos, mais le pire, c'est que mes nerfs ont complètement craqué…*which correspond semantically to the contents of the option. By concluding from the attitude shown by the speaker through almost the entire text, distractor 5c could be easily interpreted as being true. However, it is not stated in the text. There is a word-match between *province* 'countryside' in the distractor and the text, which might attract weaker test-takers.

The stem in item 6 asks what is said about the apartment. Again, all options could be true, and each has to falsified against the spoken text. Distractor 6a is a negated statement, including the modifying adverb: *trop*. These two modifying elements add to the complexity of the processing of the written part of the task (Cf. Powers 1985, Biber 1997, Freedle & Kostin 1999). The test-taker has to figure out the basic statement: "*l'appartement est cher*" 'the apartment is expensive', then transform it into the opposite: "*l'appartement n'est pas cher*" 'the apartment is not expensive'. Finally, the meaning of the polysemous and comparative adverb *trop* 'too' in this context has to be interpreted. In the text, the falseness of this statement is evidenced by the affirmative *L'achat a bien sûr demandé de gros investissements* 'The purchase naturally demanded big investments'. In the key option 6b, the word *évoque* 'evoke' might be unfamiliar to the test-takers, if they do not happen to recognize it from their English vocabulary. However, *souvenirs* 'memories' should be a familiar noun, even if it might be difficult to associate with anything in the text. The text says « *ça nous rappelle un peu la vie de province* » 'it reminds us a little of the countryside life', so the link has to be made between the nominalization of a synonymous verb (*se souvenir →souvenirs)* in the option, and a verb (*rappeler)* in the text. Another problematic issue may be the fact that this information comes before the necessary information for item 5. According to Weir (2005: 64-5), the listening test items should ask for information in the same order in which it occurs in the passage; if not, it may confuse the test takers, which could lead to unreliable performance. For the distractor 6c there is actually no corresponding text content. Nothing seems to be said about the completion of the building or restoring of the apartment. The reactions to this option will likely be confused.

Items 5 and 6 are testing, among other things, the ability to compare negated – i.e. relatively more demanding - written statements with the spoken text content, and to infer the attitudes of the speaker towards a situation, expressed mainly verbally, but also by her tone of voice. The variable mentioned by Rupp et al. (2001) concerning the directness of information and its impact on the item difficulty seems relevant in this context (See Table 9) – the requested information is implicitly provided, and inferencing is more demanding than simply recognizing information.

Items 7, 8 and 9: content analysis

This is a passage with three items related to it. However, the passage is only slightly longer than the ones combined with two items. The vocabulary contains possible problematic parts, such as: *s'installer* 'move in'*, s'inquiéter* 'worry'*, cambriolages* 'burglaries'*, exposer* 'exhibit'*, devait fermer* 'had to close'*, surveiller* 'keep an eye on'*, bagarres* 'fights'… There is a lot of rather surprising information – nothing of the contents seems to be possible to infer just by the theme or by own experience. From that point of view it is a text passage that is good to use for testing purposes: more or less everything in the text is to be understood.

Item 7 asks a very general question that does not lead us to any particular detail in the text; we have to look at the options in order to know what to focus on. The problematic trait with this test format is the fact that the textual necessary information related to item 7 comes after the information needed to respond to item 8 (cf. Weir 2005: 64-5). The introspective responses may provide information on whether this has confused the test-takers.

7a is a clear paraphrase or a summary of: *Il y a de la place, alors on a exposé des œuvres des amis peintres, des copains photographes* 'There is space, so we have exhibited pieces of art made by painter and photographer friends'. 7b demands processing of the temporal aspect of verbs in the text compared with the ones present in the option: the difference lies in the presence of present tense in the option and past tense (*passé composé*) in the text, which makes 7b a distractor. 7c looks like a good and efficient distractor, since there are textual clues in the text that seem to lead to it, for instance: *exposé des œuvres des amis peintres, des copains photographes.*

In item 8, the word *cambriolages* leads the listener to the identical word in the text. The task is to understand the text surrounding this vocabulary item, and combine it with the correct option. 8a is associated with several pieces of text – both those related to disturbing behaviour (*bagarres, cambriolages*), and those related to art. As for 8b, the opposite is clearly stated in the text: *La porte est souvent ouverte* 'The door is often open'. It can be assumed, that only weaker students fall for this distractor. Corresponding paraphrases for 8c are given: *Ce sont nos parents qui s'inquiétaient, à cause de cambriolages, et ils ne sont toujours pas tranquilles. Nous, ça ne nous a jamais fait peur. La porte est souvent ouverte.*

The word *inconvénient* 'inconvenience' in the question for item 9 is found in the text, helping the task of focusing. Some may select the key option 9a just because it can be matched with *la rue* 'the street' in the text. The sentence containing the word is the necessary information that gives the truthfulness of the option: *Les soirs de fête, je dois surveiller la rue et regarder ce qui se passe, parce qu'il peut y avoir des bagarres.* The word "inconvenient" in the stem is the main evidence against distractor 9b. The playing of the piano in the text is not seen as anything negative according to the text: *C'était sympa* 'It was nice'. In addition, the agent in the option, *voisin* 'neighbour', and the temporal adverb with an inherent element of frequency *la nuit* 'at night' do not correspond to that in the text. The inferred meaning of distractor 9c is found in the text, but with a positive connotation.

These three items 7-9 contain various elements; temporal indications, opposites and inferences are to be understood in order to handle the items successfully. Moreover, the ability to hold in short-term memory information that comes earlier in the text, to be treated with a later item is at stake as well (See the difficulty variables in Table 9 Rupp et al. 2001).

Items 10 and 11: Content analysis

In this passage, there are two different people speaking, an older and a younger woman. Item 10 is different from any of the others in that the options consist of single words that are to be used to finish the sentence of the stem. It seems reasonable to include this variation in the test. In fact, this could be said to be the only "pure" MC-question, where clearly only one of the options can be true for one situation! Here the main focus is on listening to and inferring from the text content. In fact, the key word in the text comes in the second introductory sentence. It is simply a synonym for option 10c. For those who do not catch or understand this key word, there are many "surface" hints in the text that rather leads to one of the distractors than to the key. For 10a we have words belonging to a schema connected to postal services: *des colis et du courier* 'packages and mail', *papiers* 'papers'. For 10b there is a direct word-match. A supplementary hint against the two distractors is, however, given in the stem in the form of a temporal indication: *il y a longtemps* 'a long time ago' which does not go together with the rest of the narration indicating the present tense, but with the adjective *ancienne* 'old' linked to the key word *épicérie* 'grocery shop'.

In the text related to item 11 slightly contradictory information is given. At the first two sentences the speaker seems unsatisfied with her childhood, but after that she is expressing an opposite attitude. We have *avouer* 'admit', *j'aurais bien aimé* 'I would have liked', *tout le temps quelque chose* 'always something'. Her change of attitude is expressed by *à part ça* 'apart from that'. The key 11a presents a neutral content that can be associated to *jouets* 'toys', *trottoir* 'pavement' in the text. Options 11b and 11c are both negated, increasing the processing load for the test taker. There is nothing explicitly said about friends in the text – she talks about *voisins* 'neighbours' and *commerçants* 'salesmen' that can perhaps be interpreted as such. By inferring from the text, one might arrive at selecting 11c. The modifier *jamais* 'never' is, however, rather strong compared to the text. The truthfulness of this option remains a bit unclear. It cannot be completely rejected by the text contents.

Item 10 seems to consist of a simple synonym- matching task, but there is actually more to it than meet the eye at first. The test-taker has to rely on scarce necessary information, and resist surface similarities. The options for item 11 contain modifiers that may lead strategically strong test-takers to, independently of the text, choosing the one with the weakest "*souvent*" 'often', compared to the more absolute modifiers "*pas de*" 'no' and "*jamais*" 'never'.

Summary of the content analysis for items 1-11

On the whole, the first items (1-11) seem good in many respects. Even if the interest of the text theme for the adolescents as test-takers can be questioned, the text is neutral and general, the vocabulary is not too specific or technical. It seems to lend itself well to this type of items. The questions are fairly good and varied; they are not discrete point items, but integrative in that they ask for processing on several linguistic levels: from the phonological through to syntactic, lexical and discourse processing. They demand an understanding of the main points in the text, processing of larger chunks of text, text-based inference, a combination of information from different parts of the text. As Buck says (2001: 123), longer texts tend to require discourse skills, whereas shorter texts tend to focus more on localized grammatical characteristics. Weir (2005: 74) also points out that longer texts demand more "executive resources" and more language knowledge for their cognitive processing compared with shorter texts.

It seems as if the text has been well made use of; the proportion of text and items seem optimal. Taking account of the information needed to answer the questions, there is not too much text that is redundant. However, as a trait related to the interactional authenticity (term used by Bachman & Palmer 1996) all pieces of textual information do not have to be understood in order to solve the tasks at this point of the test, which is taken to add to the authenticity of the task.

Considering the interaction between text and task further, the test-takers have to treat several different textual and syntactic characteristics: temporal indications, opposite facts, various adverbial modifiers. As a rule, the understanding of separate lexical units in the spoken text is not sufficient for answering the questions. The task for the students is also to resist superficial clues in the text – matching words in the options with identical words in the text is not a successful test-taking strategy. This speaks for the validity of these test items.

However, one problematic issue is the fact that many of the options for these eleven items represent, in fact, true/false statements, and not responses to precise questions. Clearly, this is a factor adding to the complexity of the task for the test-takers, who have to use a large amount of their cognitive capacity on processing the written text. Added to that, there may be complications due to the fact that for some passages (items 5 & 6 and 7, 8 & 9), the necessary information for solving the items do not follow the order of the item questions. For items 7, 8 and 9, moreover, three items have to be solved at once, which also adds cognitive load to the task.

The items present other traits that may influence the validity in a negative way. These mainly have to do with options that are not well formulated with respect to their relationship with the text. In order for the options to be good and transparent, the following criteria have to be met. First of all, they have to be clearly and unambiguously wrong or correct. Moreover, they should be formulated in a way that makes them easy to interpret. This is very important, since the focus should be on understanding the spoken text, and not the written. The options should therefore be short, and they should not contain unfamiliar vocabulary, expressions or syntax (see chapter 2.6.1).

Some of the items do not follow these rules, but seem to complicate the task for the test-takers, thus influencing the test task processing and, as a consequence, possibly the test outcomes in a skewed way. We may have cases of construct-irrelevant variance decreasing the validity of the conclusions made on the basis of the test results.

The first case is the presence of potentially unfamiliar vocabulary in the options. It is naturally difficult and practically impossible to see to it that all the words used in the options are always familiar to all test takers. The next best solution is to have vocabulary that is potentially familiar to all test takers on the targeted proficiency level. In this case it implies relatively frequent vocabulary, concrete vocabulary, and such vocabulary that can be met in the textbooks at use in the schools. In the current case, representing the reality in Finnish schools with a limited number of textbooks in circulation, it is possible for the test constructors to have general knowledge of the vocabulary presented. Naturally the teachers may present their own supplementary text material including a different vocabulary selection to their learners and future test-takers. It would, nevertheless, be fair if the vocabulary at use at the test would follow some principles.

It is necessary to consider the fact that it is not easy to create good options for test-takers at this ability level. The vocabulary selection is only one of the issues. As the options have to be short and scannable, synonyms and paraphrases are frequently used. From the semantic point of view, the distractors have to consist of plausible interpretations of the test content, but still be clearly wrong. There are thus limitations of several kinds that affect the creation of MC items (See references to Ebel & Frisbie 1991; Haladyna 2004, 2004; Linn & Miller 2005, chapter 2.6).

The cases with the possibly problematic vocabulary in the options are fairly difficult to foresee at the stage of item development. It partly depends on whether the unfamiliar vocabulary units are also presented in the spoken text, with possible explicit or implicit hints as to their meaning. Generally more difficult vocabulary consist of abstract concepts, sometimes those without cognates in the other language known to the test-takers such as English, Swedish or German. The following options contain that kind of abstract vocabulary: 1a) *l'attrait* ('the attraction'), 2a) *uniquement* ('only'), 2c) *liens sociaux* ('social bonds'), 3b) *(il) s'occupe* ('he takes care of'), 4b) *(ils) prennent pour* ('they take for'), 6b) *il évoque* ('it reminds'), 7c) *propriétaires* ('owners'), 8 stem: *cambriolages* ('burglaries'), 9a) *proximité* ('nearness') .

There are syntactic complexities in the use of negations and various modifiers in the options. The following are to be found: 2a) *uniquement*, 5a) *ne la dérangent pas* ('do not disturb') 5b) *vient d'avoir* ('has just had') 6a) *n'est pas* ('is not') 6c) *ne sera jamais* ('will never be'), 11 b) *n'avait pas de* ('didn't have any') 11 c) *ne… jamais* ('never')

However, the effects of the presence of these words or expressions may vary. Some test-takers are likely to spend a lot of their processing capacity trying to figure out the meaning of an unfamiliar word, leaving less capacity for the rest of the task. Some are reluctant to select an option with unfamiliar vocabulary, whereas others make use of the test-taking strategy and MC test wiseness of selecting that particular option since it is most probably the correct one. All this influences the validity of the test, since the reasons for selecting or not selecting the key option will at least to a certain degree be dependent on other circumstances than the understanding, the misunderstanding or problems with understanding the spoken text. The nature of these potential effects will be partly clarified by the test-takers' introspective responses to each of the items.

7.2 Items 25-30

The last six MC items (25-30) also have three options to choose from. There are, however, several differences in this part compared to the first part of the test. First of all, these items consist of separate text passages, with only one item per passage. The themes of the texts vary, as well as the text type. The two first ones are brief, newsflash-like texts, whereas the four last ones can be called pragmatic, in that the test-takers' task is to select the most appropriate speech line for a particular situation which consists of a short dialogue.

Item 25: Content analysis

The three key words in this text are *galleries* 'galleries', *artistes* 'artists' and *public que* 'audience'. The task for the test-taker is to find out what the main idea of the passage is. It is, however, necessary to understand practically everything, in order to arrive at the correct choice.

In 25a, it is the word match with *café* that is the probable reason to choose this option. 25b might be attractive because of the two matching keywords in the text: *accueille* 'welcome' and *artistes* 'artists'. What is more, the meaning of this option is not very unlikely considering both the text content and real-word knowledge about what a gallery might want to do. The words in the text leading to the correct option 25c are *non-initiés* ‘outsiders’ and *nouveaux type de publique* ‘new type of audience’, from which the implied meaning has to be inferred.

Item 26: Content analysis

The very short passage contains vocabulary units associated mainly with either science or an award: *Matématicien* ‘matematician’, *philosophe* ’philosopher’, *humaniste* ‘humanist’, *scientifique* ’scientific’ and *recompense ‘award’, prix* ‘price’, *palmarès* ‘top list’. Option 26a is attractive because of the presence of the (possibly) familiar historical personal name also found in the text. The weaker test-takers, who have not understood very much of the text contents, might fall for this one. 26b contains the adjective *scientifique*, also found in the text. The implicit meaning of this option is close to that of 26a. The key 26c should be fairly easy to match with the text – both *prix* and *scientifiques* are found in it.

Item 27: Content analysis

Here the main task is to recognize that the passage is a dialogue that takes place over the phone. A customer wants to talk about a specific subject, and the bank employee is asked to connect to the person in charge. Obviously, this demands declarative knowledge about how “hold on” is expressed in an idiomatic way. Consequently, even if the situation in the text is understood, the appropriate expression needs to be familiar to the test-taker. Only then can the key 27a be selected. If the test-taker is thinking logically about the meaning of the verb *laisser,* option 27b might seem a good line. Syntactically, as the verb is transitive, this expression would, however, be impossible. The verb *passer* ‘pass’ in option 27c is found in the text, and some test-takers might go for this, as it gives a “superficial vocabulary link” to the text.

Item 28: Content analysis

This text passage contains various hints that help the test-taker place the dialogue in the appropriate situation. *Partir* ‘leave’, *vacances* ‘holidays’, *TGV* (= a French express train) should lead him or her to the situation where the speakers are leaving on a train. *Places assises* ‘seats’, *reservation* ‘reservation’ are clues to the problem. By the tone of voice, the listeners should observe that there is some sort of a dispute between the two speakers. 28a is in fact the only response that goes logically with the text – it gives the reason why one of the speakers has acted the way she has. If the last statement in the passage is interpreted as a question, 28b might seem a good answer. It does not, however, match the rest of the semantic contents of the passage. 28c, on the other hand, could be a logical selection for those who ignore the pragmatic mismatch but who understand the tone of voice being annoyed, possibly indicating a quarrel between the speakers.

Item 29: Content analysis

This seems to be a fairly straightforward dialogue that takes place in a butcher's shop. The unexpected information might be the fact that the customer wants to buy something for her cat, not for herself. This seems to be the key information, along with the information on what type of meat is available in the shop. 29a could be a completely acceptable response, both grammatically and semantically, as well as pragmatically. The only objection to the choice is the fact that the customer does not buy the meat for herself. However, it still seems to be possible to respond in this way in the situation. But as 29b also seems to be a potentially correct response, the task for the test-taker is to decide which is the most probable in the context. Judgments have to be made about the speaker's character and tastes. 29c is logically impossible: the customer cannot buy something that is unavailable. For the test-takers who select their response by word matching, this might still be the chosen response.

Item 30: Content analysis

This text passage contains a speaker's suggestion about plans for the vacation, and another speaker's reaction to it. The vocabulary is not complicated, but the temporal indications might cause trouble. There are verbs in several different tenses: present tense, past tense, future, and the conditional structures of *tu pourrais* 'you could', *on s'amuserait* ' we would have fun'… The fact that the speakers are both young females makes the lines more difficult to associate with either of them (cf. Brown 1995: 59-69, here chapter 1.5.1), which is also likely to have an affect on memorization. Option 30a implies that the other speaker has to reject the suggested idea. The speaker's response indicates, however, that there is the probability that she can join them as suggested. "*Vacances*" 'holidays' in option 30b may attract weaker test-takers who rely on word-matching strategies. The line does not suit either of the speakers, as the first one talks explicitly about her vacation, and the second one implicitly. The key option 30c is a line that belongs to the first speaker, as a comment to the other speaker's reaction to her suggestion.

Summary of contents for items 25-30

These types of items represent a good complement to the more "traditional types" in items 1-11. They seem to measure a slightly different aspect of listening ability, thus resulting in a more thorough coverage of the aimed construct.

Short newsflashes assess the ability to process the message fast, as in listening to the news on the radio. Short dialogues bring the test-taker to a quasi-interactional situation, where he or she is given the task to reply, as if being in the situation. Both syntactically, logically and pragmatically acceptable answers are needed. This seems to be close to the contents of Buck's (2001) listening construct (see ch. 1.5) where only the sociolinguistic dimension would be missing. At the test-takers' current language ability level, this dimension would presumably yet be too demanding to fully be included in a test, however. Therefore, as far as the variation of the text types is concerned, these test items taken together seem to have a sufficiently good quality. Consequently construct under-representation does not seem to be the most serious problem influencing the validity of the items - except for the lack of some features of "spokenness" in the text. There are naturally always texts with language functions related to the communicative listening ability that need to be left out because of the limited time set for the test. A test is always constructed by sampling, by compromising and by judging what test items would cover different pieces of the listening ability as efficiently as possible within the practical constraints of the entire testing framework.

Among these last six items, there are a couple of potentially problematic cases that may cause construct-irrelevant variance. The first one is item 27, the traits of which have to be discussed. The item seems to be testing two things. First, it is testing comprehension of the overall situation of the dialogue. This is completely acceptable from the point of view of the listening construct. Second, it is testing the knowledge of an idiomatic expression used in that particular situation: the task is to recognize the correct expression among three written options. The acceptability of this part is questionable. One could claim that from the pragmatic point of view it is an expression that is used orally, and is thus important to be familiar with – to understand when encountered. The problem here is that even if the test-taker has solved the first subtask (understanding the dialogue), he or she cannot obtain any credit for it without solving the second (having the declarative knowledge of recognizing the written idiomatic expression). The question remains if this is valid and fair for the test-takers. We have to rely on the information provided by the item analysis and the introspective responses to find a possible answer to this.

The other problematic case is item 29. From the point of view of the content analysis and as a hypothesis of the potential judgments made by the test-takers during the processing, there seem to be two correct options: 29 a) and b). This may be very confusing for the test-takers, who should try to find evidence for and counterevidence against each of the options. There is only vague counterevidence against the correctness of option 29 a): the fact that the response may imply that the meat would be for the speaker. But an interpretation where the speaker speaks for the well-being of the cat (if she wants to buy filet for her cat the natural inference is that she cares about its well-being) that is "I do not like entrecote [for my cat]" seems completely acceptable. Subsequently some test-takers may have well understood the spoken text, but ended on a distractor that is not clearly wrong. The question is whether it is clear enough that the most suitable option is to be selected, which in this case would be option 29a.

7.3 Summary of the content analysis

To briefly sum up the content analysis, items 1-11 and 25-30 seem to cover an acceptable part of the listening construct, considering the practical limitations of the present test situation. These items are seventeen in total, and in a real test, there would be almost the double amount of items, which is naturally expected to increase the validity with respect to the fact that more text types and functions can be included. However, as was mentioned in chapter 7.1, what seems to be missing are features closer to the "actors" in the communicative situation – oral interaction where the listening feature is essential. Whether and how this component can be included in large-scale assessment contexts remains an important issue. Buck (2001) points to the necessity of making sure that the test tasks at least engage the same abilities as target language tasks – implying interactiveness (with the concept of Bachman and Palmer 1996).

The problems often related to the MC format remain a threat to the validity of the items. Possible construct-irrelevant variance may be caused by opaque options (as opposite to transparent, i.e. clear and unambiguously correct or false) containing unfamiliar vocabulary or syntactic complexities. This potentially leads to an excessive focus on the written instead of the spoken text. This situation is also present in tests of reading comprehension, as Allan (1992) has concluded in his doctoral thesis. According to his analysis of introspections, MC items engaged strategies focusing on the stem and options, compared to open responses. For a test of listening comprehension, this is evidently an even more serious problem.

Another related issue has to do with the fact that many of the options of the items represent separate statements, and each one of them has to be verified or falsified separately, since they do not rule out eachother. Exceptions to this situation are represented by the items 1, 2, 10, 26 and the four pragmatic items 27-30. The hypothetical assumptions of the effects on the validity of the interpretation of the test scores will be verified by the item analysis, and an analysis of the introspective responses of the test-takers.

8 ANALYSIS A POSTERIORI: QUANTITATIVE INFORMATION ON THE ITEMS

In order to get an idea of how the items have functioned for the current particular group of test-takers that I consider comparable with a potential target group for a test of listening comprehension of French as a L2 as part of the Matriculation Exam (see chapter 5.2.1 and the results of the chi square analysis of the original and the current test administration), I now turn to the quantitative information obtained after the items have been administered. Making sure that the items have worked in an intended way - being on a convenient level for the test-takers - is needed to be able to draw conclusions on the processes and strategies that the test-takers have employed. The quantitative information that I have taken account of and that is reported and discussed here concerns the test-takers' scores and their person measure as a function of the results in solving seventeen listening comprehension items, as well as different item values obtained by means of conducting a Rasch analysis (by means of the software *WINSTEPS* 3.69.1), analyses by means of the software *SPSS* 17.0 and distractor analyses.

8.1 The test-takers' results

First of all, to show how the test-takers as a group have succeeded in the listening comprehension items, information on the test-takers' results is presented in Figure 4 (below). For the current pool of items, where one point is earned for each correct option selection, the observed mean score is 8.4 (S.D 2.9; see Table 19 below), which implies that for an average performance, half of the seventeen items have been correct.

The exact distribution of the different observed total scores is shown in Figure 4 below. The mode is 8 (31 cases), followed by 7 (30 cases) and 10 (28 cases). The extremes, the lowest score (0) and the highest score (17), are represented by one case each.

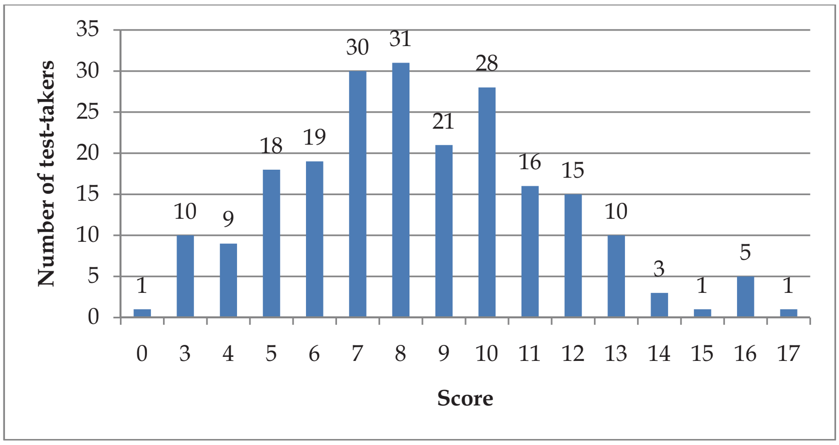


FIGURE 5 Distribution of total scores in the current test. Total number of items: 17, one point each

The Rasch model produces the person measure, with the purpose of describing the test-takers' ability in the measured skill. It needs to be pointed out that in this research project, however, this measure is not reliable for this purpose as it is based on a pool of items of a restricted size (17). In this context it can be used, for example, as a tool for comparing choices of options for the different items (as in Table 21) but it cannot provide a reliable and covering estimate of an individual test-taker's listening skill.

The mean of the test-takers' scores related to their person measure can be observed in Table 19 below. The items as a whole have been on a convenient level of difficulty for the test-takers since the means of the distribution of the person measures and the item measures meet at the value of 50 (see Figure 6 in Appendix 2). The infit and outfit values for the person measures are acceptable, so the data fit the model well. However, the reliability figures (real: .58, model: .61) as well as the separation values (real: 1.19, model 1.24) for the person measures are low. The shortness of the "test" (that is the low number of items) – added to the low number of options per item - influences the reliability of the person measure. The limited number of items do not discriminate particularly well between test-takers – the distribution of the test-takers' person measures is not spread-out.

TABLE 19 Summary of 216 measured (non-extreme) persons

	RAW	PERSON	MODEL	INFIT		OUTFIT	
	SCORE	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	8.4	50.01	5.49	1.00	.1	1.00	.1
S.D	2.9	8.86	0.87	.14	.8	.21	.8
MAX	16	79.41	10.43	1.48	3.1	1.72	3.0
MIN	3	33.39	5.07	.72	-2.1	.34	-1.9
REAL RMSE 5.71 **ADJ.SD** 6.77 **SEPARATION** 1.19 **PERSON RELIABILITY** .58							
MODEL RMSE 5.56 **ADJ.SD** 6.89 **SEPARATION** 1.24 **PERSON RELIABILITY** .61							
MAXIMUM EXTREME SCORE: 1 Persons **MINIMUM EXTREME SCORE:** 1 Persons							

8.2 Results of the Rasch analysis of the items

The Rasch analysis provides information on the fit of the items to the Rasch model. All the items are found to fit the model (item 3 having values closest to the threshold[54]). The item reliability of .94 shows a convenient level – even if this figure should not be considered in absolute terms, as the items in this research study do not constitute an entire test, but only a part of it. Tables 20 (in Appendix 2) and 21 show a summary of the Rasch analysis and a quantitative analysis of each item. As the items are found to fit the model, this provides a good basis for the introspective analysis and for looking further into the functioning of the separate items.

The item measure tells about the difficulty of the item - here the mean difficulty for all items is set at 50.00. The seventeen items that are studied vary in their difficulty measure between 63.37 for item 25 (see Table 20), being the most difficult item, and 36.45 for item 11, being the easiest one. As far as the discrimination is concerned, the lowest discrimination indexes are found with items 3 and 9.

TABLE 21 The selection of options (key/distractor) for each of the 17 listening comprehension test items related to the test-takers' person measures

Item	**Key (selection %, average person measure)**	**Distractor (selection %, average person measure)**	**Distractor (selection %, average person measure)**
1	1b: 49%, 53.49	1a: 40%, 45.39	1c: 11%, 50.09
2	2c: 55%, 53.76	2a: 17 %, 44.56	2b: 28%, 45.94
3	3c: 42%, 53.11	3a: 22%, 48.20	3b: 36%, 47.39
4	4a: 40%, 54.54	4b: 40%, 47.78	4c: 19%, 45.16
5	5b: 51%, 53.49	5a: 17%, 45.55	5c: 31%,: 46.76
6	6b: 37%, 54.55	6a: 31%, 47.95	6c: 32%, 46.71
7	7a: 56%, 52.76	7b: 20 %, 46.04	7c: 24 %, 46.90
8[55]	8c: 54%, 53.77	8a: 21%, 47.79	8b: 25%, 43.82
9	9a: 47%, 53.11	9b: 16%, 45.18	9c: 37%, 48.08
10	10c: 43%, 54.22	10a: 15%, 47.75	10b: 42%, 46.55

[54] For a high-stakes test, MNSQ values between 0.8 and 1.2, and ZSTD between –2 and 2 are generally accepted (see Linacre, J.F. http://www.rasch.org/rmt/rmt162f.htm and http://www.rasch.org/rmt/rmt83b.htm, consulted 15.5 2011.)

[55] Item 8 is the only item where two test-takers have not made any option selection at all. The figures for this item are thus based on 216 test-takers' responses.

11	11a: 76%, 51.67	11b: 11%, 44.47	11c: 12 %, 44.88
25	25c: 24%, 55.57	25a: 10%, 48.56	25b: 66%, 48.22
26	26c: 57%, 53.11	26a: 22 %, 45.01	26b: 21 %, 46.90
27	27a: 33%, 54.34	27b: 28%, 48.27	27c: 40 %, 47.67
28	28a: 65%, 53.11	28b: 25%, 44.06	28c: 10 %, 44.98
29	29b: 48 %, 54.22	29a: 48%, 46.99	29c: 5%, 37.49
30	30c: 67%, 52.56	30a: 28 %, 45.02	30b: 4%, 42.66

The average person measure of the test-takers having selected a particular option is related to the discrimination of the item. The larger the difference between the average person measure for the key and the distractors, the stronger the item separates between weak and strong test-takers (see Table 21). As an example: for item 2, with the discrimination index of 1.28, the average person measure for the key is 53.76 and for the distractors 44.56 and 45.94 respectively. For item 3, with the discrimination index of .63, the average person measure for the key is 53.11 and for the distractors 48.20 and 47.39. Thus, for item 2, compared to item 3, the difference between the mean level of success for test-takers who have found the key and those who have ended on a distractor is larger, which implies that item 2 discriminates better between these two groups.

The combination of the item measure with the discrimination index on a scatterplot (see Figure 7) gives us an idea of the relationship between these two values. What is clear is that there is no linear relationship between the variables. There are items of mid-difficulty with very different discrimination indexes, like items 2, 3, 5 and 8. There are also items which have discrimination indexes close to 1 but are of very varying difficulty, like items 5, 6, 11 and 25. They are all within the range of acceptable items, but it can be assumed that the qualitative analyses will reveal at least some of the factors that affect where exactly they are placed in the figure.

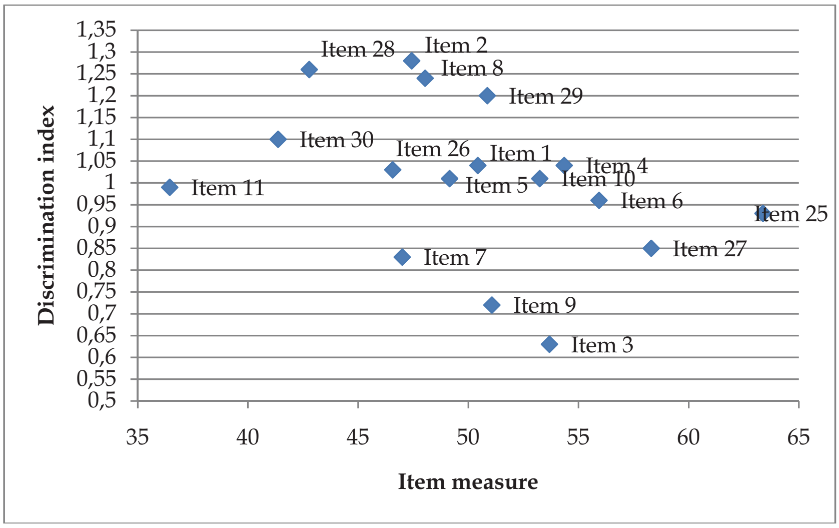


FIGURE 7 Relationship between item measure and discrimination index of 17 items

8.3 Distractor analysis

As far as the selection pattern of the options is concerned, for an item working well, both distractors should ideally attract a sufficient number of test-takers. According to Emslie & Emslie (2005: 4) an item is acceptable if the percentage of test-takers selecting each distractor is between 20/d and 80/d where d is the number of distractors. For the items in the current test this means that for a good distractor, it should thus be selected by between 10 and 40 % of the test-takers.

The items with distractors that each attract between 15 and 35 % of the test-takers are (see Table 21 above): 2, 3, 5, 6, 7, 8 and 26. There are items in this test where a distractor has been more attractive than the key (items 10, 25 and 27) or not particularly attractive (less than 5% of choices: items 29 and 30). As is pointed out by Emslie & Emslie (2005: 13) a distractor should be plausible enough to attract uninformed students and all distractors should be evenly attractive, no distractor being redundant. For all the present seventeen items, the average person measure is higher for the key than for either of the distractors – this indicates that the distractors are selected by the generally weaker test-takers (defined as their relatively weaker result on these 17 items), which is what is intended.

One way of getting a graphic summary of the option analysis is to take the percentage of choices of a key or a distractor per group of test-takers. Below a graphic analysis for item 2 (item measure: 47.4), a seemingly well-functioning item, is presented (Figure 8). In order to avoid a sawing pattern caused by a low number of test-takers for individual score groups, the test-takers are here divided into three larger groups according to their total scores: 1-6 points out of 17, 7-11 points and 12-17 points. The division of test-takers for the trace lines is as follows (Table 22)[56]:

TABLE 22 Division of test-takers into three score groups

Score	Number & proportion of test-takers
12-17	32 = 15 %
7-11	129 = 59 %
0-6	57 = 26 %
Total	218 = 100 %

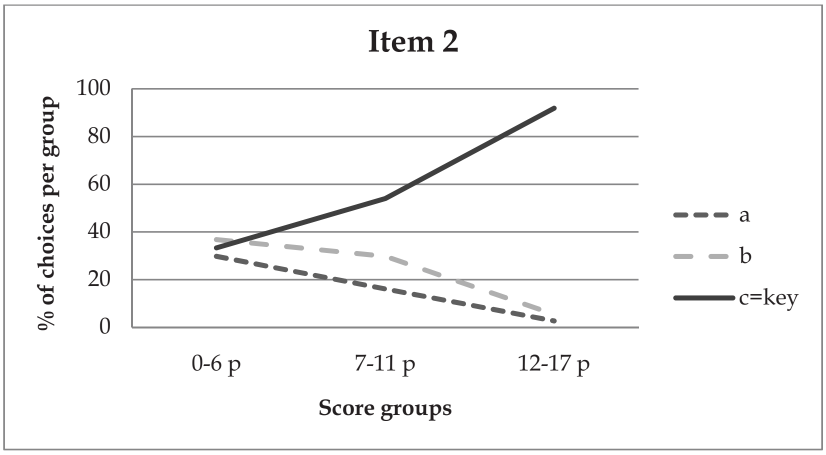


FIGURE 8 Trace lines for the selection of the key (2c) and the distractors (2a and 2b) as a function of the proportion of cases within three score groups (0-6 points, 7-11 points and 12-17 points)

For item 2, both distractors are selected by test-takers with relatively lower total scores, while the key gradually becomes more frequently selected until it dominates to over 90 % among the high-scorers. This can be taken as the targeted pattern for a well-functioning test item.

As examples and to contrast with item 2, the trace lines of three other items are presented: the easiest item, item 11, the most difficult item, item 25, and the item with the lowest discrimination, item 3. The trace lines for item 11 (item measure: 36.5) shows a similar pattern to item 2 (see Figure 9), the differ-

[56] The justification for this division is based on different facts. The border between the first and the second group is drawn at 6, since there seems to be a “jump” in the number of test-takers between the scores 6 and 7 (see Figure 4). The borderline between the two other groups rests on the idea that the 11 first items differ in their type from the 6 last ones. Thus, with a limit set at 12 points for the high ability group, it is not enough for a test-taker to have solved the 11 first items correctly; at least one of the other type of items needs to be correct also.

ence being that the trace line for the key is very high on the diagram, whereas the trace lines for the distractors remain at the bottom, which is typical for a very easy item. In fact, for item 11, already at 5 points 60 % of the test-takers have selected the key.

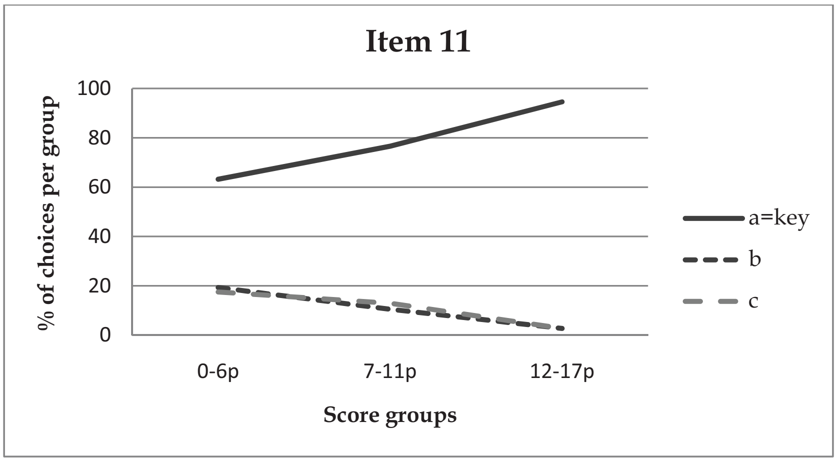


FIGURE 9 Trace lines for the selection of the key (11a) and the distractors 11b and 11c as a function of the proportion of cases within three score groups (0-6 points, 7-11 points and 12-17 points).

In the case of the difficult item 25 (item measure: 63.4), there are clearly visible differences to item 2 (see Figure 10). First, the attraction of distractor 25b shows as a line very high on the diagram: it has been fairly attractive on all score levels. Distractor 25a on the other hand, has been selected by equal proportions of test-takers within the three groups. The slope for the trace line of the key 25c is moderate. It is only at the level of a total score of 14 that the majority (more than 50 %) of the test-takers have selected the correct option (see Figure 11), compared with item 2, where this happens at a total score of 9.

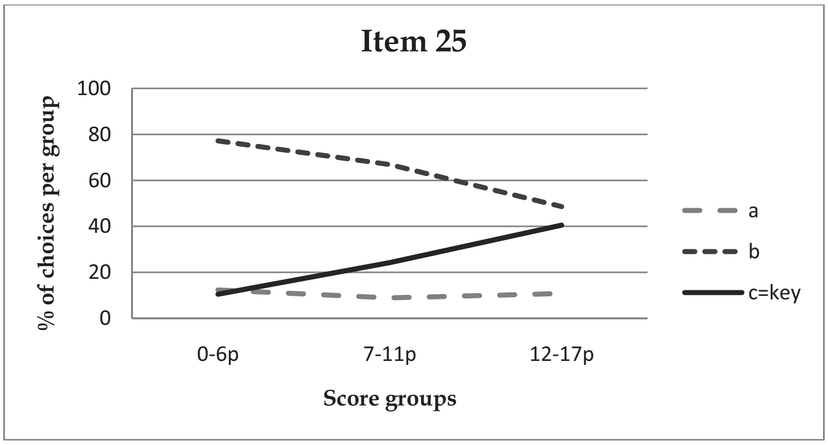


FIGURE 10 Trace lines for the selection of the key (25c) and the distractors 25a and 25b as a function of the proportion of cases within three score groups (0-6 points, 7-11 points and 12-17 points).

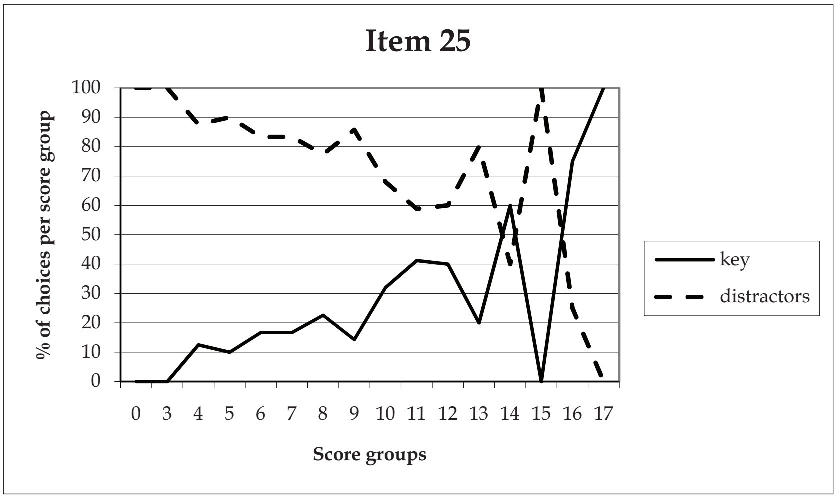


FIGURE 11 Trace lines for the selection of the key or a distractor for item 25 across the individual total score groups.

For the purpose of further comparison, the trace lines for item 3 (item measure: 53.7) are drawn in order to see whether they show any differences compared to the other items due to the relatively lower discrimination of the item. The trace line for the key reaches only 60 % for the group with the highest scores, with both the distractors being attractive across the three groups. The low discrimination is thus graphically visible here as flat trace lines across the three options.

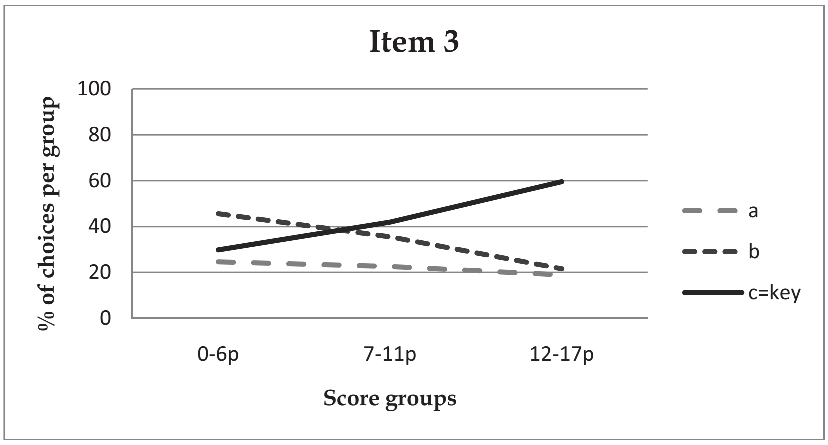


FIGURE 12 Trace lines for the selection of the key (3c) and the distractors 3a and 3b as a function of the proportion of cases within three score groups (0-6 points, 7-11 points and 12-17 points).

8.4 Comments on the quantitative analysis

The combination of different pieces of quantitative information that I have reported above can first of all be taken to justify the qualitative study that follows (chapter 9). The quantitative analysis indicates that the items have functioned according to the expectations of the Rasch model, and they can be considered acceptable items[57], even if they are showing some differences in their quantitative characteristics. More thorough information as to the functioning of the multiple-choice items is expected to be found in the qualitative analysis – details that are impossible to detect by a mere quantitative analysis. In the following chapters, an analysis of the introspective responses will help further in identifying items that have functioned well and items that have not functioned as intended. Do the test-takers' introspective responses somehow reflect the situation where there are different processes or strategies at stake when the test-takers are confronted with for example the different cases of difficult or easy items, or items with low discrimination?

[57] This is as expected, since the items are constructed and used within a high-stakes assessment context.

9 THE RESEARCH QUESTIONS WITH A FOCUS ON THE INTROSPECTIVE RESPONSES

The questions that I will treat in this chapter are: What processes and strategies are activated or employed by the test-takers at one particular MC test of listening comprehension of French as a foreign language? How are these processes and strategies related to the outcome for the individual test-takers on the seventeen items? Do the scores given on the basis of the seventeen items seem to be a valid and reliable reflection of the intended (part of the) construct? The assumption is that the short written introspective responses taken together will provide information on this matter. Thus, first all the introspective response categories and types will be presented in the following (chapters 9.1 – 9.2). In chapter 9.3, I will describe the different introspective responses and discuss them in more detail, with examples provided for different items by different test-takers. In chapter 9.4 I focus on whether and how the processes and strategies reflected in the test-takers' introspections have influenced their results on the 17 listening items. In 9.5 I will discuss what problematic features in the items are revealed by the introspective responses.

9.1 Description of different types of introspective responses

The introspective responses can be divided into four (4) larger types according to their focus. Some of them are text-focused; they mention words, idea-units or propositions, contexts or text contents. Others concentrate on the strategies related to the task. There are also responses that give comments on the test situation itself or reflect some emotions experienced by the participant. Finally, some response boxes are left unfilled, and yet other responses are so vague, that they do not provide very much information at all. In Table 23 below I present the nine different introspective response types with a brief description of each. As this format of introspection has not been used, or at least reported on, before, there is not any previous study to refer to directly. I have therefore categorised these types on the basis of the data obtained from a pilot test and the current test. Some of the types can, however, be related to earlier inventories or lists of listening or test-taking strategies.

TABLE 23 Description of the different introspective response categories

Text-based responses	
Word-bound responses	The participant gives an answer containing single, separate words from the text, or states that "this word was heard in the text".
Partial comprehension	The participant gives an answer that contains some semantic similarities with the spoken text, but which does not cover all of the needed spoken text information.
Résumé	The participant gives a more or less covering summary of the content of the text; the main information needed to be able to select the correct option and discard the distractors.
Nonsense responses	The participant gives an answer that contains an interpretation of the text that has little or no similarities with the semantic content of the spoken text
Task-related strategies	
Option-focused responses (Elimination)	The participant states that s/he has used the ruling-out strategy, or that s/he has of some reason been able to discard one of the options
Guessing	The participant states that s/he has selected his or her answer by guessing.
Comments on the "meta-level"	
Metacognitive comments	The participant gives comments on his or her thinking about the text or the task, about him- or herself in the particular test situation or about some background information or experiences.
Vague or empty responses	
Vague responses	The participant does not seem to know how to justify his or her choice of an option. S/He just states, "It was said in the text", "I just felt this" or something similar.
Empty responses	The participant leaves the answer box empty.

The introspective responses are naturally not always clear-cut and easy to place in a certain category. There are also cases where the introspective response includes a combination of two – or in rare cases many – types of responses. An example of a response, given by test-taker (N10:9) to item 5 includes both evidence of partial comprehension and option-focused elements:

- *Ei mainittu muutosta, ja äänet häiritsivät häntä → c oli ehkä lähimpänä '* They didn't mention the move, and the noices disurbed her → c was probably the closest'

Another example by test-taker (G1:3) to item 4 could, in fact, be placed in four different categories (which represents, however, a rare case, and comes from a test-taker among the eight who had already practiced with some of the current items and were therefore excluded from this study): guessing, metacognitive, word-bound and nonsense responses:

- *Arvaus, toisella kerralla kuulin sanoja turisti, yms. Häntä luullaan turistiksi, kun hän valokuvaa* ' A guess, the second time I heard words like tourist etc. People think he's a tourist when he takes pictures'

For the sake of being able to count some quantitative differences, I have determined a hierarchy of response types, with the basic principle that the text-based responses dominate the others, followed by the metacognitive responses, with the task-based responses coming last. This is motivated and justified by the fact that the text-based responses are felt to reveal the processes at the core of the comprehension of speech, referring directly to the spoken text and being an obvious element included in the construct. The metacognitive responses, on the other hand, are very heterogeneous and are expressed in very different ways by different test-takers faced with different affective reactions to different "micro"-situations within the test. However, they tell more about the test-taking situation than the simple indication of the strategy of guessing or elimination. Moreover, as guessing is given explicitly as a suggested strategy in the test instructions, the researcher assumes that there are more indications of this strategy than what is actually the case. The vague and empty responses usually represent clear cases. According to this categorisation, the first example above is thus placed in the category of partial comprehension, while the second is placed in the nonsense-category.

There is naturally an impact of the principles of this hierarchy on the results and thus the conclusions of the quantitative analysis of the introspective responses. However, the researcher believes that the tendencies provided by the quantitative results are valid and informative, especially as they are combined with a detailed qualitative analysis of the introspective responses. Therefore they give interesting information on how the processes and strategies occur as a function of the characteristics of the different items, affected also by the skill or the personality of the different test-takers.

Taken together quantitatively at first, the introspective responses provide an idea of what processes and strategies are at stake in a situation where the L2 listening comprehension ability is assessed by means of MC test items. The proportion of different types of introspective responses given as answers to all the 17 items by all the 218 test-takers is shown in Figure 13.

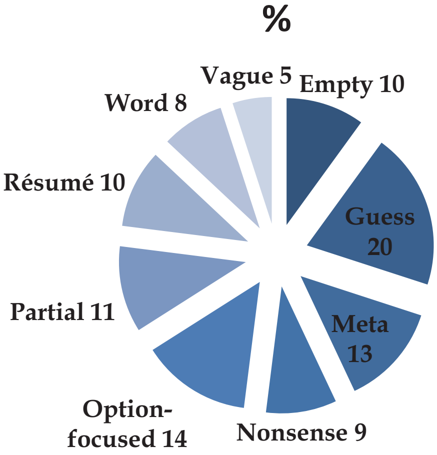


FIGURE 13 Proportion of different responses to all items by all test-takers.

It is clear that the nature and the selection of processes and strategies are different from a L2 listening situation where no test items, or other item formats are used. A graphic presentation of the hypothetical optional processes and strategies at stake in a MC test-taking situation is shown in Figure 14 (in Appendix 3). The process contains at least four stages, which, however, are not necessarily linear, but closely interrelated. We first have the context and the imposed purpose for listening – which can be paralleled with basically any listening comprehension situation in general – but which, in a MC test situation, consists of the task, including the question (stem) and the response options (See Yi'an 1998). If the questions are read through before the listening stage, as is the case in the current test situation, the test-taker can either get a preliminary idea of what is to be expected in the spoken text to come, or he or she may be confused about the situation. In either case, this is something that influences all the rest of the stages in the process, and which is an important issue to consider (See also Jamieson et al. 2000; Table 7 above).

At the stage of listening to the spoken text input, the degree – the quantity and the quality - of comprehension varies, both between each separate listening event or test-taking situation and between individual test-takers. At one extreme end, nothing is understood, whereas in the other extreme end, practically all of the text, or at least the necessary information, is understood (see the first stage in the information processing model by Jamieson et al. 2000; Table 7 above).

The reactions and actions related to the task vary depending on the degree of comprehension reached as well as on the expectations or questions raised through the preliminary task. The interpretation of the text – the mental representation of the text contents - is to be compared with each of the options (As in the second stage of the information processing model by Jamieson et al. 2000). If nothing is understood, the test-taker either needs to make a random guess on one of the options, or eliminate the ones that seem improbable based on clues other than from the spoken text. In case single words from the text are understood, the selection is based on these – which probably imply a word-match with the same concept in a distractor – or then the strategies of guessing or ruling-out are used.

Partial comprehension of the text is likely to bring the test-taker closer to the key option: this information on the text contents may be sufficient for selecting the key, especially when combined with the use of the strategy of elimination. If the test-taker considers this partial information to be insufficient for the option selection, or to base elimination on, he or she may still rely on guessing. If the main content of the text – the necessary information - is understood, there should not be any problems selecting the correct response option. However, due to for example problems in understanding the question or the options, a test-taker who has understood the text contents may still arrive at a distractor.

At the fourth stage of the test-taking situation, the likelihood of selecting the correct response option increases with the degree of spoken text that is understood (cf. stage 3 in the information processing model by Jamieson et al. 2000). However, a random guesser or a word-matcher may be lucky and arrive at a correct option, while someone that has understood the text may be distracted from the correct interpretation and select a distractor (see Figure 14).

9.2 Consistence between coders for the introspective responses to item 3

In order to establish the validity of the nine categories of introspective responses that I have selected, I let two other coders code the responses given by all 226[58] participants to one randomly selected item, item 3 (Cf. the recommendations by Bachman 2004: 279 and Green 1998: 94). The coders are both linguists and researchers of French, with Swedish as their mother tongue, high-level skills in Finnish and English, and experience in language teaching.

I gave them the descriptions of the nine categories together with a table of hierarchy[59], with the help of which they were to determine the category in cases where two categories (or more) may come into question. They both worked independently, and did not ask any questions during the task.

In 71% of the cases (161 out of the total 226 = 0.712) there was total agreement among all three coders: we had all assigned the same category to the response (see Table 24). When these cases are added to the cases where either of the two other coders agreed with the original coding, the agreement reaches 90 %

[58] Here are included the eight test-takers that were then discarded from the study as they were already familiar with the last six items (25-30) from their classroom practice.

[59] The basic principle is that the text-based responses dominate the others, followed by the metacognitive, with the task-based responses coming last. (Justification provided above in chapter 9.1)

(204/226=0.903). The cases where neither of the two judges agreed with the original coding amount to 22, or to 9.7 % of the cases. Of these, in eight cases all three coders disagree.

TABLE 24 Agreement on the coding of item 3 by three independent coders

Total agreement	71% of cases
Agreement with at least one other coder	90 % of cases
Disagreement with both other coders	10 % of cases
Coding of all three coders differ	3.5 % of cases

In several of these cases of total disagreement, it is a question of having assigned different levels of text comprehension to the introspective response. Where I have accepted the response as being a résumé, the two other judges have been less lenient, or the other way around. Examples of such responses are:

- *Tycker att han sa nånting om att han tycker om att vara ute* ' I thought he said something about being outdoors' (H1:4)[60] researcher: R[61]; judge 1: P ; judge 2: vague
- *Hän työskentelee kadulla.* (U5:6) researcher: P; judge 1: R ; judge 2: W
- *"je me crois à la plage", "entouré par des arbres et des fleurs". Ei pidä turisteista, ja ihmiset jotka eivät työskentele häiritsevät häntä* ' "je me crois à la plage", "entouré par des arbres et des fleurs". He doesn't like tourists, and people who don't work bother him' (J13:11) researcher: P; judge 1: R ; judge 2: W

Other responses are interpreted as being misunderstandings by the researcher, while the judges have considered them being in accordance with the text. In two of the responses there are, in fact, both true and untrue elements that the judges have estimated as being more or less pertinent:

- *Hän työskentelee maalarina kadulla* ' He works as a painter in the street' (K2:3) researcher: NON; judge 1: R ; judge 2: P
- *Hän tykkää olla ulkona ja maalata* 'He likes being outdoors and paint' (F3:5) researcher: NON; judge 1: R ; judge 2: P

In another response, the test-taker has interpreted "ne travaillent pas" in the text as corresponding to the notion of unemployment, present in the option selected:

- *Hän puhuu jotain työttömyydestä* 'He says something about unemployment' (Z9:5, B)[62] researcher: NON; judge 1: W ; judge 2: P

Finally, there are two cases where I have seen the metacognitive elements as being the pertinent contents of the introspective response, where the two other judges have not:

- *Hän puhui kai työstään, joten vaihdoin vastauksen* ' I think he was talking about his job, so I changed my response' (S1:2, A→B)[63] researcher: META; judge 1: P ; judge 2: W

[60] Test-taker (H1:4), option selection: key option

[61] R= résumé; P= partial comprehension; W= word-bound response; NON= nonsense; OPT= option-focused response (elimination); G= guess; META= metacognitive response;

[62] Test-taker (Z9:5), option selection: a distractor

[63] Test-taker (S1:2), a change of option choice to a distractor.

- *Ei taidettukaan sanoa, että ne olisi nimenomaan turisteja* ' I think after all they didn't say that they were tourists in particular' (F4:7, A→B) researcher: META; judge 1: W ; judge 2: ?

The metacognitive category is the most heterogeneous, with meanings on different levels, so the categorisation of this response type seems to demand more reflection than the other responses. All these examples of different judgments illustrate the challenges of the task and the need of sufficient time and familiarity with the different types of responses to be able to be consistent and precise. It has indeed often been necessary to read through each individual response several times, comparing it with the spoken text and the task.

9.3 A closer investigation of the different introspective response categories

As an assumption, the different introspective responses are given as a function of the nature of the item as well as of the test-takers' characteristics and is related to the test-takers' success level. In the present chapter I will describe the introspective response categories further with examples of the different responses provided. This will reveal more about the nature of the processes and the strategies that are activated when the test-takers are solving the listening comprehension test items. I will first consider the different text-based responses (word-bound, partial comprehension, résumé and nonsense), and then present the two strategies of guessing and elimination[64]. Finally, I will describe the heterogeneous category of meta-cognitive responses.

9.3.1 Text-based responses: Word-bound responses

Some test-takers, especially less successful ones, have only managed to grasp separate words in the spoken text, and are obliged to form an interpretation of the spoken text contents on the basis of these words in relation to the suggested options (Cf. "less active –listener" strategies and "word-oriented" strategies mentioned by Nakatari 2006). There is a significant correlation between the person measure and the frequency of the use of word-bound responses. The tendency is that the lower the person measure, the more frequent is the use of the word-bound response type (See Table 25 in Appendix 4 and chapter 9.4 below). From the point of view of the test-takers' processes activated in the test situation and from the point of view of the characteristics of separate items, there are issues to consider further related to this type of responses. It is informative to

[64] As far as the two introspective response categories guesses and elimination are concerned, for these closer analyses the cases differ slightly from the analysis where all the introspective categories are taken into consideration. In other words, here all the cases that can be interpreted as containing some element of guessing or elimination are treated, even if they may belong to some other category in the analysis as a whole (according to the categorisation hierarchy, see footnote 40 and ch. 9.1 and 9.2).

see what kinds of words the test-takers have picked out. Another question is whether some items incite more word-based responses than others. And if so, is this related to the estimated item measure or the discrimination of the seventeen items in the test?

The proportion of word-bound responses for the different items

As far as the number and proportion of word-bound responses is concerned, for some items they have been particularly frequent, whereas there are a few items where there are hardly any cases at all (See Figure 15).

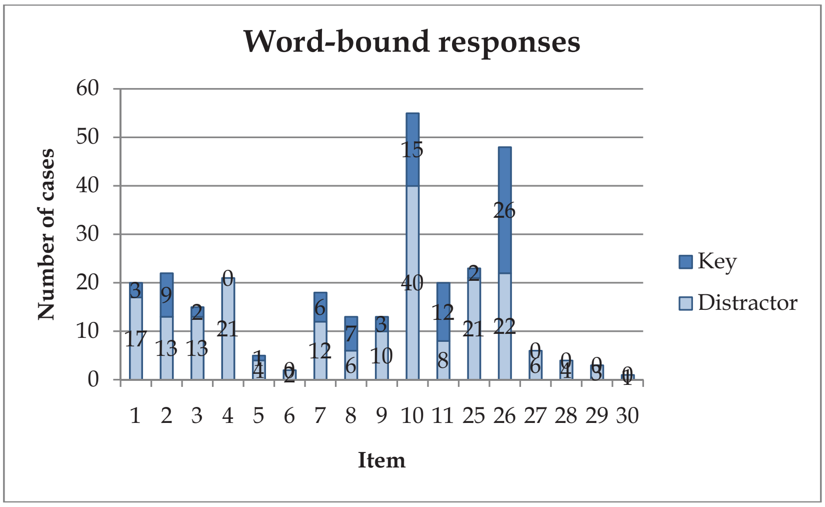


FIGURE 15 Number of word-bound responses and success per item

Items 10 and 26, where the options consist of single words or concepts, seem to naturally call for more reliance on single words in the text. The word-bound responses are thereby particularly frequent for these items (16 % of the total number of word-bound responses are given to item 10 and 17 % to item 26).

There are other items that only cover for 2 % or less of these responses: items 5 and 6 (which refer to the same spoken text passage) and items 27, 28, 29 and 30 (being the "pragmatic" items, where the options represent a missing line). For the last four items, this phenomenon seems to be explained by the fact that the options do not directly target to paraphrase the meaning in the text, but they represent a logic and reasonable continuation to the presented dialogue. The word-matching strategy is simply not fruitful.

For items 5 and 6 there does not seem to be a clear superficial or external explanation as to the reason for the low number of word-bound responses. The two items seem to discriminate well and be of mid-difficulty. However, looking closer at the introspective response patterns, it can be noticed that for item 5 there are dominantly other text-based responses than word-bound responses - that is responses reflecting partial comprehension or in some cases summaries of the text contents. For item 6, on the other hand, there is a generally very low total number of responses based on the spoken text, but a very large number of cases where the response is based on guessing. The conclusion is that whereas the test-takers have generally grasped more than just single words from the text in the passage for solving item 5, for item 6 they have not frequently been able to find, or at least refer to, any element in the text as a basis for their option selection. They have therefore been forced to make a guess, due to a difficult text, problematic options or possibly an interaction of several factors[65].

The total number of word-bound responses for these items amounts to 289. Of these, only 86 cases (approximately 30 %) are given in combination with a correct choice of an option. This indicates that the reliance on separate words is generally not a good strategy for selecting an option.

As another observation on the relationship between the quantitative characteristics of the items and the word-bound responses, a calculation of the Pearson correlation shows that there is no significant correlation between the item measure and the number of word-bound responses (See Table 26 in Appendix 4). This means that there are other factors than the difficulty of the item explaining the behaviour of the test-takers faced with the individual items. The interplay between the text, the items and the individual test-taker's way of reacting and processing is naturally decisive – as exemplified by the difference between items 5 and 6 above. It is thus necessary to take account of the entire response pattern for an individual item in order to arrive at a possible explanation for the test-takers' behaviour and at other information related to the quality of the item.

Information of a more qualitative nature enlightens the general character of this type of responses. All the 289 cases of word-bound responses can be divided into five subtypes according to the following:

- The test-taker has written down a high frequency word appearing in the text (but **not** in any of the options), or a hyponym / synonym (abbreviated as **WT** in the following);
- Two high frequency words from text are given **(WT2)**;
- The test-taker has given one or two words that appear both in the text & the options, either in the key or in a distractor **(WO)**;
- A "metacomment" is given **(WM);**
- The test-taker mentions words that do not appear in the text or in the options **(WNT).**

Two subtypes of responses (WT and WT2) thus focus on words in the spoken text that do not appear – at least not in the same shape – in any of the distractors or the key. The test-taker has given either one or two single words directly from the text, or synonyms or hyponyms covering words in the text. The subtype with **one single word from the text (WT)** is found most frequently with item 10 – and with half of these responses combined with the selection of the key option. In these correct cases, a key word in the text has been sufficient to

[65] Item 6 is discussed further below, in chapter 9.5.

reach the correct interpretation. Item 10 is different from the other items in that the options consist of only one word or concept – the task being to match one word in the text with a synonymous expression in the key.

Examples of these responses given for item 10 are:

- *Puhuja mainitsi jotakin paperista…'* The speaker mentioned something about a paper…' (J8:6, A)[66]
- *Puhuttiin jostain myynnistä...* 'They talked about some selling…'(Å2:4,B→C)[67]
- *Puhu jotain vastaanotosta…*'They said something about a reception…'(S3:4, B→A)
- *Ravintolaan viittavaa sanastoa* 'Vocabulary referring to the restaurant' (Å4:5, B)
- *Sana épicérie viittaa kai siihen* 'The word épicérie probably refers to that' (N8:7)[68]
- *Une vieille epicerie* (S12:7)

This subtype is also frequent for item 11, where 'loneliness' and 'pavement' are two typically focused concepts:

- *Yksinäisyyteen viittaavat sanat* 'Words referring to loneliness' (E3:3, B)
- *Puhui yksinäisyydestä* 'Talked about loneliness' (E4:4, B)
- *'trottoir' mainittiin* 'trottoir' was mentioned' (K16:7)
- *Mainittiin jalkakäytävä* 'The pavement was mentioned' (B11:8)

Slightly more than 40 % of the total number of responses of the subtype WT (given for all items) are on the key – the success depending on what specific words have been grasped in the text.

If **two words from the text (WT2)** have been combined as a basis for selecting an option, the results reflect that the probability of finding the key seems greater compared with the situation where only one word is used: more than 50 % of the cases where two words from the text are quoted are on the key option. This type has been particularly frequent for item 26, where the clear majority of the cases of WT2 are combined with the selection of the key. As is the case for item 10, the options in item 26 also consist of single concepts – not entire clauses – which seem to call for more focus on single words in the text:

- *Puhuttiin filosofiasta ja René Descartesista* 'They talked about philosophy and René Descartes' (X2:5, A)
- *Descartes ja palkinto ja tiede, ei ole löydetty tai huomattu mitään.* 'Descartes and prize and science, nothing has been found or noticed' (F3:5)
- *Joku eurooppalainen palkinto* 'Some European prize'(K17:8)
- *Kuului "un prix européenne"* '"Un prix européenne" was heard ' (U9:8)

The most typical subtype among the word-bound responses is the one where one or two **words appearing in the options (WO)** are given as a response. The processing that seems to lie behind these responses reflects an unsuccessful strategy to base the selection of an option on. The reason is that words that are found both in the text and the options are words that the test constructors typically use for creating distractors: a word from the text in a completely different context. This is also indicated by the low success rate: only 17 % of the responses of this subtype (WO) are associated with a correct response.

[66] = Test-taker J8:6, option selection A, a distractor

[67] = Test-taker Å2:4, first option selection B, a change to C, a distractor.

[68] = Test-taker N8:7, option selection: key.

This subtype occurs in combination with all the items, but is especially frequent for the two items where one single word or concept is given as an option. In item 10, the word that the test-takers have relied on is *restaurant:*

- *Kuulin sanan ravintola ja arvasin* 'I heard the word restaurant and I made a guess' (Item 10: T3:4, B)

For item 26, it is the proper name *Descartes* that appears in the majority of these cases of WO and this has lead to a distractor:

- *Låter som Descartes* 'Sounds like Descartes' (D1:2, A)
- *Descartesista puhuttiin paljon* 'They talked a lot about Descartes' (U5:6, A)

However, the test-takers who have grasped the key word *prix* in item 26 have found the correct option:

- *Palkinnoista puhuttiin nauhalla* 'They talked about prizes on the tape' (X4:6)
- *Puhuttiin palkinnosta...* 'They talked about prizes…'(Å2:4)

In item 4, different versions of the frequent concept of *photographe* 'photography' *photographier* 'take pictures' can be found in the word-bound responses, as well as of *touriste* 'tourist' or *touristique* 'tourist-like' or their combination. All these response processes have lead to a distractor:

- *Puhuttiin valokuvaamisesta* 'They talked about taking pictures' (J2:2, B)
- *För att jag uppfattade ordet touristique* 'Because I grasped the word touristique' (H1:4, C)
- *Valokuva sana mainittiin* 'The word photography was mentioned' (P8:5, B)
- *Turisteista ja valokuvista puhuttiin*. 'They talked about tourists and photographies' O6:5, C→B)

There are a few responses of the subtype **WM** where the test-taker seems to have reflected on his or her own basis for selecting an option:

- *Perustuu joihinkin kuultuihin sanoihin* 'Based on some heard words' (Item 8*:* E3:3, B)
- *Sanat mainittiin* 'The words were mentioned' (Item 9: J5:4, B)
- *Kuulin nauhalta vastaukseeni sopivia sanoja* 'I heard words on the tape that matched my response' (Item 10: R1:2, A)

Interesting cases are those where the test-takers have "heard" words that are not in the text **(WNT).** At item 2, some test-takers seem to have drawn further conclusions based on single words, combining a word that is in the text with another one that is not. Their conclusions do not correspond to the intended text interpretation. However, the test-takers who have grasped the word *social* in the text have also selected the key, as the key includes a word-match with this word in the text:

- *Puhui jostain taloudellisesta vaikeuksista* 'They talked about some financial difficulties' (F1:1, C→A)
- *Arvasin. Puhuttiin "sosiaalisesta ongelmasta*" 'I made a guess. They talked about "a social problem" (X2:5)
- *Puhuttiin kalliista vuokrista* 'They talked about expensive rents' (B4: 4, B)
- *Puhuttiin sosiaalisuudesta ja asiakkaista* 'They talked about sociality and clients' (U3:5, B→)

Item 7 also provides examples of this type of responses where the test-taker has "heard" words that are actually not in the text, but are in some cases in the options. Their interpretation has been shaped according to these words, which then has given the basis for option selection:

- *Työ yhdistettynä kotiin* 'Work combined with home'(K2:3, C)
- *Tauluista yms. puhuttiin* 'They talked about paintings etc.'(K5:3, C)
- *Mies puhuu galleriasta* 'The man talks about a gallery' (Z9:5, C)
- *Siitä taidenäyttelystä oli juttua* 'They talked about an art exhibition' (K20:9)

The tendency is that the type of word-bound response is related to the test-takers' scores for the task of solving the seventeen test items. For the type of response where a word occurring in both the options and the text is referred to (WO), it is more frequent among the test-takers with a lower score than with test-takers with higher scores. The situation is the opposite for one or two words from the text that do not appear in the options (the types WT and WT2): these types are more frequent among the test-takers with higher scores compared with test-takers with lower scores. This lets the researcher assume that there may also be a similar relationship between the test-taker's skill and the use of the word-bound responses: weaker test-takers tend to use it more than stronger test-takers. The test-taker's level of success is thus related to both the number (or percentage) of cases of word-bound responses in general (the higher the scores, the less there are of this type of responses altogether) and the subtype of word-bound responses.

There are some general conclusions on the test-takers' processing to be drawn from the analysis of the word-bound responses. The tendency seems to be that the listening comprehension process of the weaker participants is not consistently automatic enough for them to be able to handle both the text and the questions and their interaction. They have not, however, given up trying to sort out what single content words there are in the spoken text that they might understand. In most cases, they show the behaviour pattern of attempting to match single words presented in the options with their spoken equivalences. This leads in the majority of cases to the selection of a distractor. In a few cases, however, the items are constructed in a way that what is demanded is an ability to distinguish a single word in the spoken text – typically not a very "foreign language learner-frequent" word – and match it with the correct synonym or definition presented among the options.

9.3.2 Text-based responses: Partial comprehension

In many cases, test-takers have grasped and understood some parts of the spoken text, without getting at all the necessary information corresponding to the main text contents usually needed to select the key option at an item. These types of responses provide interesting information on the way the text is processed: what part of the text seems most prominent to the test-takers? Why is it not or why is it sometimes sufficient for selecting the key at an item? Does the number and nature of these responses reflect the characteristics of individual test items?

As far as the differences of the test-takers' responses according to their success in solving the items are concerned, there is a significant correlation between the person measure and the number of partial comprehension responses (see Table 25 in Appendix 4): the higher the test-taker's person measure, the more frequent is his or her introspective response indicating partial comprehension.

There are three subtypes of answers where partial comprehension is evidenced. First, the most common type is the one containing idea units that appear both in the text and in the options. Second, some test-takers give responses indicating partial comprehension with one or two idea units from the text that do not appear in the options. Third, there are also test-takers who present idea units from the text but combine these with information that is not in the text, through misunderstanding or wrong inferences.

In investigating the quantitative relationship between the individual items and the number of responses of partial comprehension, it can be determined that there is no significant correlation between the total number of responses of the partial comprehension-type and the item measure (See 26 in Appendix 4). This has to be considered together with the fact that the number of responses with partial comprehension can be lower due to a greater number of word-bound responses, nonsense responses or responses that are not text-related at all – which would be the expected case for a more difficult item. The other explanation is that there are fewer cases of partial comprehension because the résumé-responses or the successful cases of elimination are frequent – which would be expected for an easier item.

Partial comprehension and success per item

For seven items (1, 2, 4, 5, 9, 11 and 29), the number of responses reflecting partial comprehension exceeds the mean (24). As for the correct responses among these cases, for some items it seems that partial comprehension has been sufficient in order for the test-takers to arrive at a correct option selection. For items 5, 7, 8, 11, 26, (27)[69], 28 and 30 more than half of the partial comprehension responses are on the key option. A combination of a large total number of partial comprehension and a large proportion of correct responses among these is only found with items 5 and 11. Of these, item 11 has been very easy, whereas item 5 is of mid-difficulty.

[69] Item 27 cannot be taken into account here due to the extremely low total number of cases.

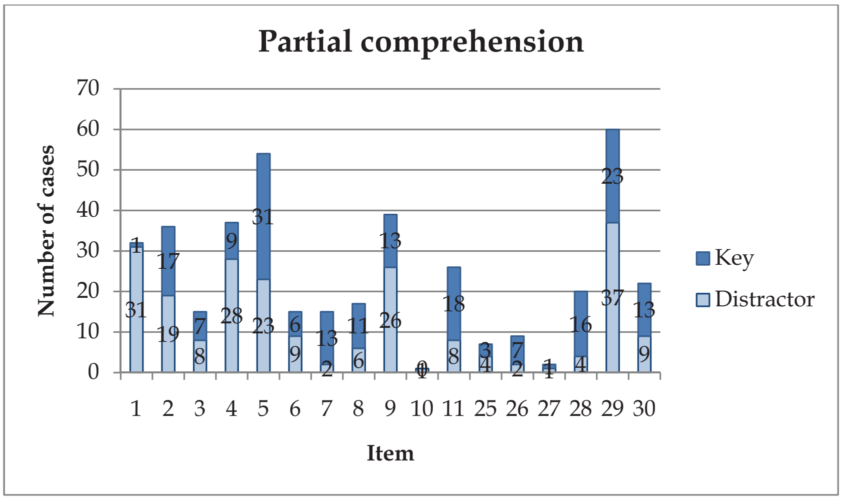


FIGURE 16 Number of partial comprehension and success per item

The many cases of partial comprehension for item 5 are related to the nature of the necessary information that consists of several pieces of information that the key option sums up. However, not all of these pieces or clues may be necessary in order to arrive at the key – thus understanding a part of the necessary information has in many cases turned out to be sufficient.[70] Examples of partial comprehension include:

- *Nainen oli laihtunut* 'The woman had lost weight' (B4: 4)
- *Melu häiritsi nukkumista* 'The noice disturbed her sleep' (K14:6)
- *Puhuttiin painon laskusta* → *terveysongelma* 'They talked about the weight-loss → health problem' (B11:8)
- *Puhui melusta, laihtui monta kiloa* 'Talked about the noice, lost several kilos' (N12:11)

This seems to give an idea of the nature of this type of responses: if elements in the text are interpreted correctly, even if they are not in the heart of the necessary, key information, it may be sufficient for selecting the key option – in some cases with the help of guessing or elimination. Skilful test-takers have probably made efficient use of strategies like inferencing and elaboration on the basis of their partial comprehension (Cf. analysis conducted by Young 1997, and below on the combination on partial comprehension and elimination).

In investigating the reasons behind the results for the individual items, we have to verify the characteristics of the responses to these items. The exceptionally low success rate among the responses reflecting partial comprehension for item 1 (with 3 % correct only) can partly be explained by the fact that as a first item in a new test, it is bound to be complicated - which is also reflected in the large number of changes from one option to another in this item (see Table 32 in

[70] This is also a question of the way the researcher has defined the borderline between partial comprehension and résumés: the résumé-responses for this item demands the mentioning of several pieces of information (the weight-loss, the annoying noices as well as the problems with the nerves), not only one or two.

chapter 9.3.7). However, the test-takers have, at the point of writing down their introspective explanations, listened to the text twice, thus having had the chance to form a picture of the text contents as a whole. The nature of the statements that the options for item 1 consist of may be decisive. As for the key option, the vocabulary may be complicated (*l'attrait* 'the attraction'; *logements* 'apartments'), making this option less appealing. The distractors, on the other hand, contain statements that are actually partly true according to the text – they just do not answer the question of the item that implies that the MAIN phenomenon has to be identified. As option 1a, a distractor, contains concrete information that is in fact expressed explicitly in the text, it is understandable that the majority of the responses witnessing partial comprehension (24 responses out of 32) refer to that piece of information, and obviously, go together with the selection of that option:

- *Tekstissä puhuttiin pienten kauppojen lopettamisesta*. 'In the text they talked about small shops closing down' (V2:3, A) (Person measure: 40.50)
- *Jag tycker att jag uppfattade som att dom sa att små butiker stängs* 'I thought I grasped that they said that small shops close down' (H5:6, A) (Person measure: 53.89)
- *Puhuttiin pienistä kaupoista ja niiden kiinni menemisestä* 'They talked about small shops and the fact that they close down' (B11:8, C→A) (Person measure: 59.51)

It is only among the strongest participants that the partial comprehension is inclined towards the correct response: elements supporting the correct interpretation of the main necessary information are given in the response.

- *Tavalliset asunnot ovat tulleet liian kalleiksi* 'Ordinary appartments have become too expensive' (Å12:7) (Person measure: 58.74)
- *Puhuttiin pariisilaisten halusta korttelielämään* 'They talked about the Parisians' wish for local life' (N11:9, A→) (Person measure: 62.76)

Contrary to this case, for items 7 and 28 the success rates among the responses evidencing partial comprehension are high – above 80 %. For item 7, the contents of the introspective responses vary a lot. Judging by these responses, there is not one identical idea unit (or piece of information) that the test-takers have focused, but various pieces of information have called upon the test-takers' attention. The fact "the door is open" has been frequently quoted, by test-takers with generally lower person measures, even if it does not seem to have a direct connection to the options for item 7. It does have a connection to item 8, however, where the opposite is claimed in option 8b. This shows the situation where three items (7, 8 and 9) linked to one text passage, and all the nine options within these items, influence the interpretation of the text and the test-taking process.

Other facts that are referred to in the introspective responses for item 7 concern the ownership of a bar, which has probably been taken as a fact to rule out the two distractors that mention other places (an art gallery and a kindergarten). The fact that the speaker has children or daughters or that she is involved with children has been referred to in five responses.

- *Asuvat baarissa* 'They live an a bar' (O15:8)
- *Koska tarha meni kiinni, he ovat antaneet ystävien/lasten olla heillä* 'Because the kindergarten was closed, they let friends/children stay at their place' (O9:6)
- *Ils vivent avec leur petite fille* 'They live with their little daughter' (N9:8)

In item 28 there are many pieces of information to focus on in the text in order to select the appropriate option, and many test-takers have referred to some of this information. The key has generally been found if the test-taker has understood that one of the speakers (the man) asks the co-speaker (the woman) if she didn't know or points out that a reservation is compulsory:

- *Junassa pitää olla varaus* 'In the train you have to have a reservation' (X3:5, B)
- *Sopii parhaiten koska mies kysyi jotain että etkö tiedä että se on pakollista TGV-junassa* 'This is the most appropriate one, since the man asks something like don't you know it's compulsory?' (X5:6)
- *Han frågade om hon inte visste att man måste reservera en biljett* 'He asked if she didn't know that you have to book a ticket' (H12:10)

It is interesting to note that there are not very many cases of guessing with this type of evidenced processing – apparently the test-takers have felt relatively confident with the pieces of information they have grasped, which implies that they have not felt the need to guess[71]. Altogether only eleven examples are found, among which there are the following:

- *Arvaus, pieniä kauppoja on jouduttu sulkemaan* ' A guess, small shops have had to be closed down' (Item 1: P10:6, A)
- *Arvasin, sillä nainen puhuu vain siitä kuinka alun jälkeen nyt on mennyt paremmin.* ' I made a guess, since the woman only talks about how after the beginning evrything has now been better' (Item 6: O9:6)
- *Gissning, talade något om ny publik...?*'A guess, they said something about a new audience...?' (Item 25: H8:7, B)
- *Ei puhuttu vain yhdestä, vaan monista herran saavutuksista...kai tämäkin oli arvaus* 'They didn't talk about just one, but many of the gentleman's achievements...I guess this was a guess also' (Item 26: U8:7, A→B)

There are a few examples where partial comprehension is combined with elimination[72]. The responses consist mainly of comments on the truthfulness of the statements in the options as judged by the partial comprehension of the text contents:

- *M. on laihtunut. A ja C tuntuivat huonoilta vaihtoehdoilta* 'M has lost weight. A and C seemed like bad options' (Item 5: U9:8)
- *He omistavat baarin, niin muut vaihtoehdot ei käy* 'They own a bar, so the other options are not good' (Item 7: K18:8)
- *"Etkö tiennyt sen olevan pakoll.?" → B ja C ei käy* ' "Didn't you know it's compulsory?" → B and C are not good' (Item 28: N9:8)
- *Koska lomanvietosta sovittiin. A on sopiva luonnehdinta. C ei liity keskusteluun* 'Because they made plans for the holiday. A is a suitable charaterization. C doesn't relate to the conversation' (Item 30: S11:6, A)

Hesitation or uncertainty is reflected from many of the responses. They seem to indicate either that the test-takers are not certain that they have understood the piece of text correctly, or that their interpretations do not fully match one of the

[71] For further analysis of the strategy of guessing, see chapter 9.3.5

[72] For further analysis of the strategy of elimination, see chapter 9.3.6

options. In the latter case, either their interpretation of the text is correct but not sufficient, or then they have misinterpreted the spoken text, sometimes led astray by a distractor. Among the examples containing an indication of hesitation, eight responses combine with the key option, whereas nine responses combine with a distractor. Examples of these are[73]:

- *Prata om turister som tog fotografier?* ’Talked about tourists who took pictures?’ (Item 4: H4:6, C)
- *Hän puhui terveysongelmista. Lihoamisesta?* ’She talked about health problems. Gaining weight? (Item 5: V2:3)
- *Sanottiin, että uudessa asunnossa on kaikki hyvin, vaikkei se ole vielä valmis?* ’They said that in her new appartment everything is ok, even if it’s not ready yet?’ (Item 6: V2:3, C)
- *Ei ollut taiteesta puhetta, ja lapsiakin vain yksi, kai* ’There was nothing said about art, and there was only one child, I think’ (Item 7: Z12:6)
- *Vieraat luulee taloa oikeesti baariksi? kö?* ’Guests actually think the house is a real bar? Or? (Item 9: U1:3, A→C)
- *Puhuttiin vissiin jotain et ihmiset tulee kyselemään ravintolasta?* ’I think they said something about people coming to ask for a restaurant ?’ (Item 10: Z7:4, B)

An interesting issue that is related to the implications of the fact that the test-takers have understood some parts of the spoken text correctly, and selected an option accordingly, is the possibility of giving partial scores (issue also discussed in Rantanen 2003: 69ff) This means that the options would be considered more incorrect or less incorrect, so that one distractor that is not completely false would yield for instance half a point, compared to the key that would give one point, and the completely false distractor would yield zero (or 0, 1 and 2 points respectively). Somehow this would be fair, since there is a fundamental practical and philosophical difference between understanding a text partially and understanding nothing of it. Naturally, with this type of scoring, the nature of the question and the options would have to be reconsidered, and the scoring principles should be explicated to the test-takers.. The consequences of this type of a basis for scoring for multiple-choice items, taking account of the implications from the point of view of validity, reliability and authenticity/interactivity would be an interesting follow-up study.

As a conclusion on the information provided by introspective responses reflecting partial comprehension, we can say that this category is most likely found with participants with an “intermediate” level of success. The hypothesis is that some of the text processing for these test-takers is automatic, but more detailed or complicated relations between the different idea units are not necessarily understood, which would, usually, be a prerequisite for arriving at the correct option. In this category, there are various extents of influence from the contents of the options on the interpretation – and this is probably paralleled with the strategy of “selective attention” (mentioned for example by O’Malley & Chamot 1990 and Goh 2002) on the basis of the options. Partial comprehension most likely leads to a selection of a distractor that could contain correct

[73] Elements showing hesitation underlined by the researcher.

elements, but that is still wrong when considering the entire semantic content of the spoken text.

9.3.3 Text-based responses: Résumé

The résumé-responses are very varying: some of the participants are content with giving only the main key phrase or key content in the text. Others retell nearly the entire story, and others still take the content of the options into account as well. The outcome in 94 % of the cases is a successful selection among the proposed options, which is as expected (See Figure 17). This is how the situation should be in a valid and reliable test of listening comprehension: if a test-taker understands the spoken text, he or she should not have any trouble selecting the correct response.

An aspect to consider relates to the differences in the frequency of cases of résumé as a function of the success level of the test-taker. A hypothesis is that the majority of the résumé responses is given among the stronger test-takers and come with the selection of the correct option. A Pearson correlation of the person measure with the number of résumé-responses points to the positive correlation: the higher the person measure, the more frequent is the résumé as a type of introspective response (see Table 25 in Appendix 4).

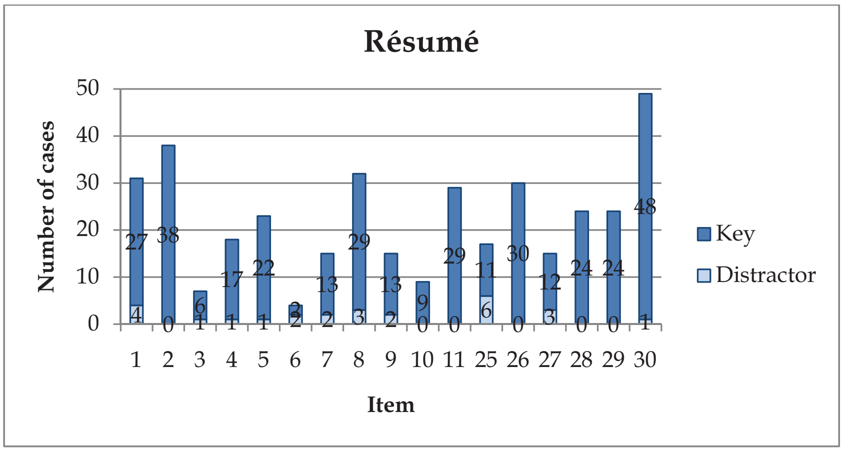


FIGURE 17 Number of résumé responses and success per item

There are cases, however, where a seemingly correct interpretation of the text has, for some reason, still lead to a distractor. The résumés that are NOT combined with a key option are interesting to look at more closely in order to analyse why this has occurred.

For item 1, four responses reflect a comprehension of the key information in the spoken text with the selection still made on one of the distractors – some hesitation has even occurred between the two distractors it seems. The responses themselves do not reveal the reason why the key option has not been selected. One possible explanation could be some problems with understanding this option, as is reflected in some of the other responses to this item (see for example the strategy of guessing, type 6, described below in chapter 9.3.5.)

- *Ihmiset ovat muuttaneet asumaan entisiin kauppatiloihin* 'People have moved to live in former shops' (S2:3, A)
- *Puhuttiin katutasolla asumisesta.* 'They talked about living on the street-level' (O6:5, A→C)
- *Pienet kaupat muutetaan asunnoiksi* 'Small shops are transformed into apartments' (U5:6, A)
- *Ihmiset muuttavat asumaan tiloihin, jotka ovat ennen toimineet kauppoina.* 'People move to live in premises that have served as shops before'(Y4:8, C→A)

Item 3 has been a generally difficult item, and the selection of the correct option has been complicated, even in the cases where the responses reflect that the test-takers have understood the necessary information in the text:

- *Hän työskentelee paikalla* 'He works there'(O13:7, B)
- *Hän työskentelee kadulla* 'He works in the street' (U5:6, A)
- *Työskentelee kadulla* 'He works in the street' (Å17:10, B→A)

In item 4, one résumé-response combines with a distractor. It seems as if the option has been misunderstood or misread – the idiomatic expression "*prendre pour un photographe*" 'take someone for a photographer' may have been confounded with "*prendre un photographie*" 'take a photo':

- *Monet valokuvaavat, häntä pidetään outona, se häiritsee* 'Many people take pictures, he is considered odd, that is disturbing' (E8:7, B)

In item 6, among the very few (four) cases of résumé-responses, two of them go exactly along the spoken text, even if distractor 6a is selected.

- *Se tuo muistoja mieleen entisestä asuinpaikasta* 'It brings up memories from the previous dwelling (K14:6, B→A)
- *Vaati investointeja, muttei vain rahallisia. Vaikutti ainoalta sopivalta* 'Demanded investments, but not only financial ones. Seemed the only suitable one' (Å11:7, A)

In item 7, a test-taker with a relatively high person measure (53.89) has made a mistake in selecting the distractor, in which a more concrete relationship is made between the fact quoted from the text and the option, compared with the key response, for which an abstraction has to be made.

- *Ystävän valokuvia esillä.* 'Showing a friend's photographs' (Y3:8, A→C)

The seemingly correct information picked from the spoken text and given in the introspective responses for item 8 has lead to different option selections. The erroneous choice of distractor 8a is made by three test-takers. In two of the wrong responses, the choice seems to be based on an incorrect elimination of the two other options, but on the basis of correctly understood test contents:

- *Ovi oli usein auki, vanhemmat pelkäsivät aluksi, kai* 'The door was often open, and the parents were afraid in the beginning, I guess' (J11:8, A)
- *Eivät pelkää ja ovi on auki* 'They are not afraid and the door is open' (R7:5, C→A)

For the third case, the test-taker seems to have used inference on the basis of a combination of two separate facts from the text:

- *Puhuttiin maalauksista ja siitä, että vanhemmat pelkäävät varkaita (arvasin)* 'They talked about the paintings and about the fact that the parents were afraid of thieves (I made a guess)' (F3:5, B→A)

In item 9, there is a similar case of the test-taker having added correct information from the text, but having made an incorrect inference on the basis of that information:

- *Arvaus, Cyrilin täytyy katsoa mitä kadulla tapahtuu. Jotkut tulevat heidän luokse koska luulevat sitä baariksi?* '*A guess.* Cyril has to see what is happening in the street. Some people come to them because they think it's a bar?' (S12:7, A→C)

The highest number and proportion of selections of a distractor combined with a résumé-response is found for item 25. Here as many as six test-takers with a résumé-response have arrived at distractor 25b. The misinterpretations are on the options or on the text and a choice of the targeted object for the event – a new audience or new artists:

- *B, koska puhuttiin uudesta yleisöstä, eikä paikka ollut kahvila*. 'B, because they talked about a new audience, and the place wasn't a café' (Q3:7, C→B)
- *Se on kaikille avoin oleva näyttely* 'It's an exhibition open to everybody' (P3:4, A→B)
- *Halunnut erikoistua, uudelle yleisölle, B luontevin minusta* 'Has wanted to specialize, for a new audience, B seems the most suitable to me' (S8:6, B)
- *Se tuo uudenlaista yleisöä uusien artistien eteen* 'It brings a new type of audience to see new artists' (Å17:10,C→B)

There are altogether relatively few cases of résumé-responses given for item 27. Two of the total 14 cases are on distractor 27b. The test-takers have understood the essential idea of the text, but have probably not known how 'Hold on' is expressed in an idiomatic manner in French:

- *Älkää laskeko luuria*. 'Don't put down the receiver' [litterarly] (P4:4, B)
- *Puhelinkeskustelu "älkää sulkeko luuria"* 'A phone conversation. "Don't put down the receiver"'(Å17:10, B)

Item 30 has the highest number of résumé-responses (49 cases) as well as a very high proportion of correct responses among these (98 %). Only one has for some reason lead to an incorrect choice of options. Could it have been caused by problems of understanding the temporal and conditional verb form in the option *aurait été* 'would have been'? :

- *Hieno idea. Täytyy ensin puhua kotona, milloin on loma jne*. 'A great idea. She'll first have to talk about it at home, when she has her holidays etc.'(O13:7, A)

One question is whether there is a relationship between the number of résumé responses and the item measure. A conducted Pearson correlation shows a significant correlation between the number of résumé responses for each item with the item measure (see Table 26 in Appendix 4). This would imply that the easier the item, the more frequent are the résumé responses. The frequency of the giving of summaries of the spoken text contents seems to be related to the functioning of an item. Thereby, items with a number of résumé responses that exceeds the average number (22) are generally well-functioning items according to the quantitative analysis: 1, 2, 5, 8, 11, 26, 28, 29 and 30. Of these, items 11 and 30 are also very easy, and items 29 and 30 have an inefficient distractor.

As we have seen, the "résumé"- responses are quantitatively very varying, as are the elements included in the necessary information for the individual items. A successful selective attention and grouping probably lie behind this type of response (See O'Malley & Chamot 1990; Young 1997 and Goh 2002). The outcome in 90 % of the cases is a successful selection among the proposed options.

9.3.4 Text-based responses: Nonsense

The giving of nonsense responses – interpretations of the text with little or no similarities with the text contents – demands the ability to create scenerios on the basis of the available clues to the text content. The weakest participants may not possess this ability or may not be capable of using it, if the entire processing capacity is either set on trying to figure out meanings of single words, or on the experienced (sometimes rather desperate) feelings in the test situation. This response type is probably most frequent among test-takers with "intermediate" scores. There is no significant correlation between the person measure and the number of nonsense responses, which may suggest that this is the case: this is neither a typical strategy for the strongest nor the weakest test-takers, but can be observed to be used among test-takers falling between these two groups.

Moreover, neither very difficult nor very easy items are clearly associated with this type of response (See Table 26 in Appendix 4). For eight items, the proportion of test-takers who have selected a correct option even if they have given a nonsense response exceeds the mean (33 %): 1, 2, 5, 7, 10, 11, 29 and 30. This can partly be explained by the facility of the items – seven of these items have lower item measures. However, the correlation analysis does not indicate a significant correlation between the proportion of correct responses and the item measure.

A closer investigation of the nonsense responses reveals some of the reasons behind the success combined with a wrong interpretation of the text. Items where both the number of nonsense responses and the proportion of correct choices among these are relatively high are items 2, 5, 7, 10 and 11. These are looked at more closely in the following.

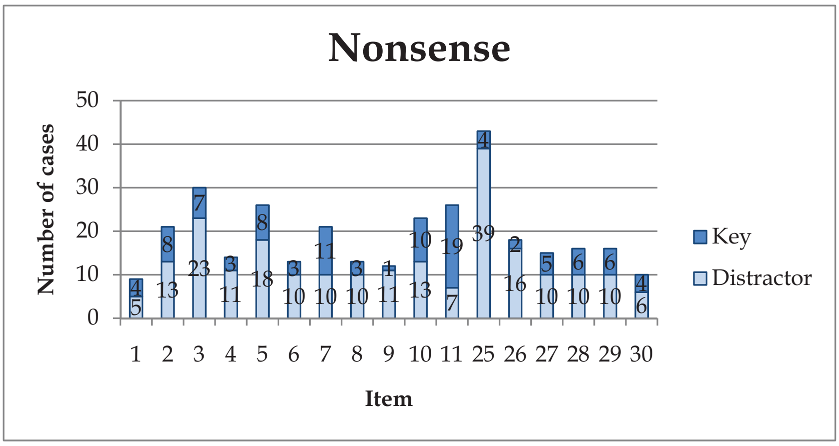


FIGURE 18 Number of nonsense responses and success per item

For item 2, with a total of 21 nonsense responses, of which eight are combined with the key, there are test-takers who have taken the social aspect mentioned in the key option to be related to a position in society. These interpretations seem to be influenced by the slight misunderstanding of the option as well as of the text :

- *Halus suurempiin piireihin sosiaalisten suhteiden takia* 'Wanted to get into broader circles because of social relationships' (T2:2, A→)
- *Olin kuulevinani, että nykyaikana ihmiset etsii "sosiaalisesti hyviä" asuntoja* 'I thought I heard that nowadays people look for "socially good" appartments' (V7:5)
- *Man berätta om människor och deras behov av en ny social ställning* 'They told about people and their need of a new, social position' (H4:6)

Other test-takers talk about the countryside, mentioned in the text, but in a context different from their interpretation:

- *Puhuttiin uudesta elämäntavasta, sosiaalisesta toiminnasta, kaupat maalla säilyisivät* 'They talked about a new way of living, a social activity, the shops in the countryside would be preserved' (S8:6)
- *Tekstissä mainitaan, että ihmiset haluavat hakeutua takaisin yhteisölliseen kulttuuriin muuttamalla pois kaupungista* 'They mention in the text that people want to get back to a new collective culture by moving away from town' (O9:6, B→)
- *Man har förutom för att man inte har råd med lägenhet vill man inte flytta ti landet* 'Other than not being able to afford an apartment people don't want to move to the countryside' (H9:8,B→)

The idea of a change is focused in two responses:

- *Ihmisiä ei enää kiinnosta myyntityö* 'People are not interested in working in sales anymore'(R6:5, B→)
- *Keskusta ei sovi enää.* 'The town centre is not good anymore' (S11:6, B→)

In item 5, where eight of the total 26 nonsense responses are on the key, many test-takers have confounded the two concepts of gaining or losing weight. However, both may be interpreted to fit into the scenery of having health problems, and thus go with the key option:

- *Melu häiritsee ja hän on alkanut lihota. Ylipaino-ongelmat = terveysongelmia* 'The noice disturbs and she has started to gain weight. Problems with gaining weight = health problems' (K12:6)
- *Äänet häiritsivät, ./. liikakilosta ja muusta puhuttiin kai, puhuttiin mielestäni Provencesta* 'The noices did disturb / I guess they talked about overweight and stuff, I thought they talked about Provence' (B9:8)
- *Lihoi paljon* 'Gained a lot of weight' (K17:8)

Also the background reasons to the health problems or weight loss have sometimes lead to confusions:

- *Menetti monta kiloa sairauden takia* 'Lost many kilos due to an illness' (P6:4)
- *Hän oli laihtunut, koska öisin liian meluisaa ja levotonta* 'She had lost weight, because it was too noisy and restless in the night' (T8:5)
- *Arvaus, hänellä oli allergia?* 'A guess, she had an allergy?' (N11:9)
- *Hon blev deprimerad av allting* 'She became depressed because of everything' (H11:10)

Item 7 also incites both a relatively large number of nonsense responses (21) and a high proportion of correct responses among these (11 = 52 %). In some cases, there does not seem to be any relationship with the text nor the correct option:

- *Lehti C & C yhteinen toive...* 'A newspaper is C & C:s common wish…'(Z3:3)
- *He asuvat samassa paikassa* 'They live in the same place'(P5:3, B→)

In some responses is reflected the obscurity of what the role of the friends that are mentioned in the key response and the text is:

- *Det hade varit en barnträdgård tidigare, de talade om vänner* 'It had been a kindergarten before, they talked about friends' (H7:7)
- *He pyytävät ystäviään omalle pihalleen* 'They asked their friends to come to their own garden/yard' (E7:6)
- *Ei aina rauhallista, koska ystäviä mm. valokuvaamassa* ' Not always pieceful, because of friends taking photos for instance' (S13:7, C→)
- *Hän antoi ystäviensä valokuvata ym.* 'He let his friends take pictures etc.'(Q5:8)

In item 10, with 24 nonsense responses, of which ten are combined with the key, some of the test-takers who have found the key have missed the temporal aspect – the fact that the apartment was a former grocery store:

- *Pieni kauppa, jonka kaikki tuntevat* 'A small shop that is known by everybody" (P6:4)
- *Puhuttiin ihmisistä ja kaupan tunnettavuudesta.* 'They talked about people and how well-known the shop is' (S11:6, A→)
- *Jotain siellä kai myytiin*…'I think they did sell something there…'(Q4:8)

Other test-takers have not understood the meaning of the key word in the options, *épicerie,* but have interpreted the text according to a misinterpretation:

- *Jonkinlaisessa "tehtaassa"* 'In some sort of "factory"'(U3:5, B→)
- *Hän oli yrittäjä.* 'She was an entrepreneur' (B6:7)

The French word *épicerie* is sometimes taken to mean a bakery, probably a mix-up with *boulangerie:*

- *Se oli ollut jokin pieni leipä-/pullakauppa* 'It had been some sort of a small bakery shop'(E9:7)
- *Hänen asuntonsa oli ennen leipomo tai vastaava* 'Her apartment was a bakery or something similar before'(N11:9)

Various other types of responses combined with the key option are:

- *Har aldrig köpt något → butik* 'Has never bought anything → a shop' (H6:6)
- *Hän pitää sisustuksesta* 'She likes the furnishing' K14:6)
- *Halusi maalaismaisemaa* 'She wanted a bit of country landscape' (Å12:7)

The item with the largest proportion of correct option choices among the nonsense responses, 73 %, item 11 also has the lowest item measure. It seems to have been easy partly due to inefficient and implausible distractors: even if only fragments of the text have been understood, some details have been misunderstood or incorrect inferences have been made, all clues still seem to have lead to the key option and to the rejection of the distractors, as is exemplified in the following responses:

- *Äiti pyysi häntä usein tekemään kaikenlaista, mutta leikki mieluiten kadulla koiranpentunsa kanssa* 'Mother often asked her to do something, but she rather played in the street with her puppy' (K10:6)
- *Arvasin. Sitä paitsi sanottiin vain ettei se asu yhdessä äitinsä kanssa. Ei mitään siitä näkeekö hän äitiään koskaan* 'I made a guess. What is more, they only said that she doesn't live with her mother, nothing about if she ever sees her mother' (U6:7)
- *Hän oli usein ulkona ystäviensä kanssa, joita hänellä oli paljon* 'She was often out with her friends, of whom she had a lot' (Å9:7)
- *Hon hade inget annat problem än att hon ibland var arg på sin mamma för att hon var enda barnet. Annars lycklig barndom.* 'She didn't have any other problems than the fact that she was sometimes mad at her mother for being the only child. Otherwise she had a happy childhood' (H12:11)

One source of misinterpretation is found in the polysemous *"jouer"* 'to play' that some test-takers have interpreted as being connected to playing music:

- *Hänen äiti vaati aina häneltä jotain, kadulla soittaessaan tyttö löytää rauhan* 'Her mother always demanded something of her, playing in the street the girl finds peace' (X1:3)
- *Hän vietti lapsuutensa soittamalla välillä kadulla äitinsä kanssa, ja piti sitä hauskana* 'She spent her childhood sometimes playing music in the street with her mother, and found it fun' (Å6:6)
- *Hän oli muiden ihmisten keskipisteenä soittaessaan* 'She was at the center of the attention of other people when she was playing music' (Å12:7)

Frequent misinterpretations (nonsense responses) are also found with items 3 and 25, where, however, the proportion of correct responses has been low - especially for item 25: 9.5 %. These items have been generally difficult and problematic; the item measures are of 53.7 and 63.4 respectively, with relatively low discrimination indexes. The misinterpretations in these items seem to be built up by the grasping of different relatively frequent concepts in the text, often combined with the contents of one of the distractors.

For item 3, there are various misinterpretations or misunderstandings based on the grasping of different pieces of spoken text – all being concrete nouns: *théâtre, trottoir, fleurs, plantes, table* (probably mixed with *tableau*). These are possibly combined with knowledge about the character of the famous arrondissement in Paris, Montmartre.

- *Taiteilja "maan tasolla", ei asu siellä* 'An artist "down-to-earth", does not live there' (F2:3, A)
- *Ei ole paljon muita myyjiä ympärillä* 'Not many other merchants around'(Å3:4, A)
- *Kiertelee kaduilla* 'Walks around in the streets' (O9:6, A)
- *Puhui työstään kukkakauppiaana, tapaa turisteja, oltava mukava* 'Talked about his work as a florist, meets tourists, has to be nice' (S8:6, A)
- *Myy tauluja ja saa työssään jutella ihmisten kanssa* 'Sells paintings and gets the chance to talk with people at his work' (Y3:8, A)'

Many test-takers seem to have taken the more or less direct meaning of option 3a and interpreted the spoken text accordingly. This results in a contrary interpretation compared with the text contents about the speaker's attitude towards talking to others:

- *Pitää erilaisten ihmisten kanssa keskustelusta* 'Likes to talk to different people' (X1:3, A)
- *Hän juttelee turistien kanssa...* 'He talks to the tourists…' (Å2:4, A)
- *Det finns mycket turister på området. Han tycker det är trevligt att diskutera* 'There are lots of tourists in the area. He likes to discuss things' (H6:6, A)
- *Turistit kysyivät häneltä saiko hänelle puhua ja hän piti heitä miellyttävinä* 'The tourists asked him if they could talk to him and he thought they were nice' (O14:8, A)

For item 25, distractor 25b has been very attractive, and is selected by a larger number of test-takers than the key. The majority of the selections are based on a misunderstanding of the text contents. There is clearly a scale of severity among the cases of misunderstanding. The interpretations that are the most far away from the spoken text are:

- *Hon verkade vara nervös för ett uppträdande* 'She seemed to be nervous about a performance' (H7:7, B)
- *Siellä voi helposti esiintyä vaikka ei uskaltaisikaan* 'It is easy to perform there even if you wouldn't dare' (N11:9, B)

Other test-takers have also interpreted "artists" not as painters and creators of pieces of art, but as performing artists:

- *Ihmisiä käy paljon ja siellä esiintyy artisteja.* 'A lot of people come by and there are artists performing' (Å1:4, C→B)
- *Puhui esiintyjistä* 'They talked about performers' (U3:5, B)
- *Siellä käy myös tunnettuja ihmisiä tavallisen yleisön joukossa.* 'There are also famous people in the ordinary audience' (O11:7, C→B)
- *Se tuo uudenlaista yleisöä uusien artistien eteen* 'It brings a new audience to see new artists' (Å17:10, C→B*)*

There are test-takers who interpret the presence of an artist as something special for the Rayon Vert, thus not understanding that it is an art gallery, where there are supposed to be exhibitions of artists' pieces of art:

- *Rayon Vert sai hieman kuuluisuutta artisteilta* 'Rayon Vert become a little bit famous thanks to artists' (R9:6, B)

- *Sanoo: tunnettujen ihmisten avulla saadaan paremmin asiakkaita* 'She says: with the help of famous people we can get more clients' (Z9:5, B)

Many test-takers believe that le Rayon Vert is a coffee shop:

- *Kahvila on esitellyt monia uusia artisteja yleisölle* 'The coffee shop has introduced many new artists to the audience' (Z5:4, B)
- *Hänen kahvilassaan on esillä uusien taiteilijoittenkin töitä* 'There are exhibited pieces of art of new artists in her coffee shop' (K10:6, B)
- *Hän sanoi tapaavansa artisteja kahviloissa* 'She said she meets artists in the coffee shops' (E9:7, A→B)
- *Artister kommer på kaffe när de går förbi på vägen* 'Artists passing by in the street drop in for a coffee' (H11:10, B)
- *Kahvila haluaa tuoda vähemmän tunnettuja taiteilijoita tunnetuiksi* 'The coffee shop wants to make less famous artists famous' (N4:6, B)

The main source of confusion concerns the target group of the gallery: the artists or the audience. There is probably an influence from the distractor in the following cases:

- *Galleria, jossa näytteillä tuntemattomampia töitä* 'A gallery where more unknown pieces of art are exhibited' (O4:3, B)
- *Taidegalleriasta puhuttiin sekä taiteilijoista jotka ovat katsojille uusia* 'They talked about an art gallery and artists that are new for the audience' (E4:4, B)
- *Paikalla on esillä vähemmän tunettujen taiteilijoiden töitä.* 'Less famous artists' pieces of work are exhibited there' (Å5:5, B)
- *Se yrittää herättää ihmisten mielenkiinnon uusiin artisteihin.* 'It tries to raise people's interest for new artists' (Å2:4, B)
- *Antaa uusia taidevinkkejä yleisölle.* 'Gives new tips of art for the audience' (Y3:8, B)

The fact that these nonsense responses are given mostly in combination with a choice of distractors shows that they work in a "valid" and reliable way, since a misinterpretation is not expected to lead to an earned score. However, there are probably many cases of misunderstanding also among the responses where the test-takers have indicated that they have made a guess. Among these are probably also correct choices.

Some general conclusions about this type of responses can be made. First, the nonsense-responses show how the test-takers have built sceneries and schemata based on not only what they have caught in the text, be it larger chunks of the text or separate words, but also combined this information with the information provided in one, or many, of the options or in another item.

Second, these types of responses are not automatically associated with more difficult or less difficult items. This is due to the fact that the proportion of nonsense responses has to be considered together with the combination of all the different types of responses at an item. A well-functioning item can have a relatively large proportion of nonsense-responses (as item 2), because a focus on the text seems to be a sign of a better item compared to one where many test-takers have relied on random guessing, for example. If an item is far too difficult for the target group of test-takers, there may be a relatively large amount of misinterpretations, even among stronger test-takers. On the other hand, an item that is too easy, due to for instance implausible distractors, can yield cases where the misinterpretations have still frequently led to the key response.

Third, this tells about the nature of the MC item: no matter what the quantity and quality of the comprehension of the text is, the key option can in many cases be reached by means of "invalid" strategies. The key option may be obviously correct to the test-taker, compared with the distractors that are eliminated, even in case of a deficient comprehension of the text. Or, a random lucky guess may save a misinterpretation of the text (See further in the following).

9.3.5 Strategy-based responses: Guessing

Perhaps the harshest and the most justified criticism on the assessment of the listening comprehension ability by means of the MC format concerns the test-taker's possibility of using the strategy of random guessing to solve the test tasks and thus to arrive at a correct response and a score without understanding the spoken text at all.

In order to obtain more detailed information on the strategy of guessing occurring in this test of listening, I have selected among all the introspective responses those where the test-takers have included some variant of the verb "to guess" or the noun "a guess", in their mother tongues (Finnish: *arvasin, arvoin, arvaus, veikkaus;* or Swedish: *gissade, gissning*). This yields 977 responses in all to focus on, which amounts to 26 % of all responses that the 218 test-takers have given to seventeen items. Thus I have taken into account not only the responses categorised as guesses in the study as a whole (responses with nothing else than the noun/verb, representing 20 % of all responses), but also responses that reveal something more about the test-taking process and are actually placed under some other introspective response category in the study (text-based or metacognitive).

At a selection-test like the MC test, some amount of guessing will probably always be included among the strategies applied by the test-takers. It can be claimed that any response where the test-taker has been uncertain of her comprehension of the text or of the relationship between the text and the options contains the element of guessing to some degree. Uncertainty, to different degrees, is reflected from many of the introspective responses in this study (see for example chapter 9.3.7).

The interesting points to investigate here are, first of all, **how** the strategy of guessing is used and, second, **why** the test-takers use it. These two questions are interrelated and partly dependent on each other. Guessing is not always, or even rarely, completely random, point that is also expressed by Bachman & Palmer (1996: 204-205) who say that there is a big difference between random guessing and informed guessing, where the latter implies that a test taker can narrow down the number of possible correct responses by using partial knowledge of the spoken text.

The situations where a test-taker relies on guessing do not just reveal something about the general nature of the test format, but also - and possibly to a greater extent - about the quality of a particular test and of particular test items. We have to consider the possibility of construct-irrelevant variance, which implies that factors outside the construct being measured affect the outcome of the items, thereby decreasing its validity. Guessing thereby has to do with both the reliability and the validity of the test scores. The fact that much guessing has occurred distorts the reliability of the scores and implies that the test-takers' responses are based on guessing due to some characteristics other than something included in the construct. This gives a test that measures something else than the construct and produces unreliable results for both the individual test-taker and for the test administration as a whole.

In this study, the majority (739 cases, ranging from 21 to 94 per item) of all the responses that include the element of guessing consists of just the noun or the verb, not giving away anything else about the nature of the strategy. The number of correct choices among these guesses is 290, which amounts to 39 % of the cases. In 238 cases the test-takers have given some kind of explanation of why or how they have guessed. On the basis of these responses, I have established ten different subtypes of guessing altogether. It is fair to assume that the 739 guesses without any indications of a reason for the use of the guessing strategy could also belong to one of these subtypes. In the following, the ten subtypes of guessing are presented and exemplified. The frequency and the proportion of correctness among these types is shown in Figure 19 below.

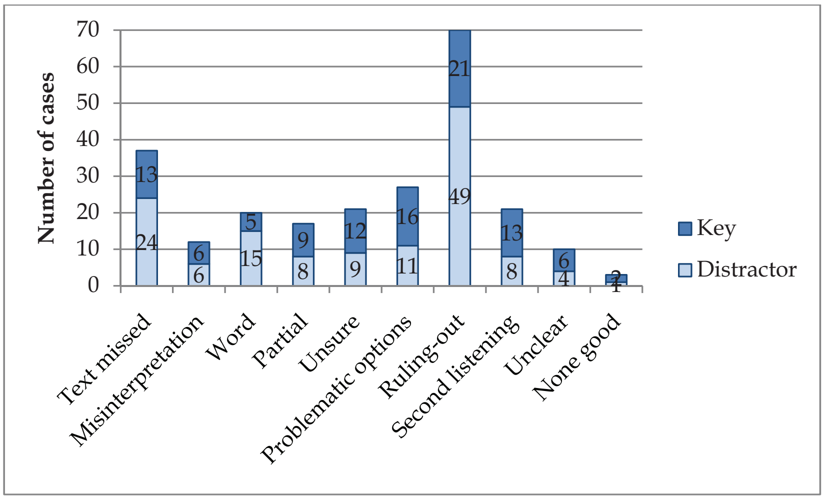


FIGURE 19 Number of cases of the ten types of guessing

Type 1 guessing: Missing the spoken input

The cases representing a non-understanding or a missing of the spoken input (due to problems in segmenting, parsing, recognizing, grasping or understanding)[74] and resulting in more or less completely random guessing could be said

[74] Difficulties mentioned by Dickinson (1987) and Goh (2000); Table 2 in chapter 1.3.

to be acceptable from the point of view of test validity. A problem from the point of view of test score reliability is, however, the cases where this uninformed test-taker happens to select the key option (problem also evidenced in the study by Yi'an 1998). A case of construct-irrelevant variance occurs if this guessing or selection is not random, but is based for example on some information in the written options giving away the correct response. In an authentic language use situation the listener can in most cases use the interlocutor or some factor other than the basic auditory information to cover for the lack of catching a message immediately. However, if the options give themselves away by their wording and test-takers can rely exclusively on test wiseness, the test is obviously not assessing comprehension of the heard message.

37 test-takers have given non-understanding as a reason for guessing; 35 % of these test-takers have chosen the key. Examples of the first type of guessing include:

- *Pelkkä arvaus, kuuntelu meni ohi* 'Just a guess, I missed the spoken text' (Item 3: E2:2, B)
- *Arvaus, puhui liian nopeasti ja epäselvästi* 'A guess. Talked too fast and unclearly' (Item 3: X3:5)
- *En oikein ymmärtänyt mitään, joten arvasin* 'I hardly understood anything, so I made a guess' (Item 5: S1:2)
- *En saanut selvää, arvaus* 'I couldn't make it out, a guess' (Item 10: T1:1, B→A)
- *Arvaus, ei ymmärtänyt mitään* 'A guess, I didn't understand anything' (Item 11: B12:8)
- *Arvasin, en oikein saanut selvää puheesta* 'A guess, I couldn't really make out the speech' (Item 25: R1:2, B)
- *Puhuivat liian nopeasti, vastaus on puhdas arvaus* 'They talked too fast, the answer is a pure guess' (Item 28: J1:2, A→C)

Type 2 guessing: Misinterpretation of the spoken text

A few test-takers (12) have guessed on the basis of a misinterpretation of the text contents. If a test works optimally, such misinterpretations should lead to the selection of a distractor, which is the case with half of these responses. In two cases, the misinterpretations match the contents of the key or at least appear to be based on the key option. This situation can hardly be completely avoided when the multiple-choice format is used, even in cases where the quality of the options is good.

Item 27 is different in that it is based on a dialogue where the missing line of speech consists of one of the options. It is thus a question of a misinterpretation of the text and consequently an erroneous inference of the situation. In some cases a guess may lead to a correct choice:

- *Arvaus, nainen kai lähti ja pyytää miestä odottamaan kunnes palaa* 'A guess, the woman probably left and asks the man to wait until she comes back' (Item 27: T4:4, B→)
- *Menee etsimään viestiä (?!). Arvaus…*'Goes to look for the message (?!). A guess…'(Item 27: Y3:8, C→)

Examples from other items include:

- *Kaupungissa ihmiset luovat enemmän suhteita kuin maalla (osittain arvaus)* 'In a town people establish more relationships than in the country (partly a guess)' (Item 2: K14:6, C→B)
- *Arvaus, hänellä oli allergia?* 'A guess, she had an allergy?' (Item 5: N11:9)

- *Se maksu kai aleni tai jotain, joten arvaan A:n* 'The fee went down or something, so I'm guessing A' (Item 6: T1:1, A)
- *Arvasin. Sitä paitsi sanottiin vain ettei se asu yhdessä äitinsä kanssa. Ei mitään siitä näkeekö hän äitiään koskaan* 'I made a guess. What is more, they only said that she doesn't live with her mother, nothing about if she ever sees her mother' (Item 11: U6:7, A)
- *Nainen lomailee perheensä kanssa. (Arvaus)* 'The woman is on holiday with her family. (A guess)' (Item 30: Å1:4, A)

Type 3 guessing: Single-word comprehension[75]

The following type of responses includes the cases where the test-takers have caught one or two single words from the spoken input. They have, however, found this limited comprehension insufficient to base an interpretation of the input and the selection of an option on, and have consequently made use of the strategy of guessing. We have in all 20 cases of this type among the responses to the seventeen items under scrutiny. It is evident that we have already moved some way away from random guessing. The selection is an individual "best guess" of some kind, based on incomplete comprehension but not being completely random. Unfortunately this has not proven a good strategy for the test-taker: only one fourth of the test-takers have selected the key. This follows the general pattern for the cases where word-bound responses are given. (In fact, these responses are categorised as word-bound responses in the study as a whole.) From the point of view of the test constructors the test probably works as intended, since a typical technique for constructing distractors is to pick a word from the text and place it in the wrong context to attract weak test-takers. The following examples show this type of response:

- *Puhuttiin turisteista ja valokuvista? osittain arvasin* 'They talked about tourists and photographs? Partly a guess' (Item 4: V2:3, B)
- *Puhu jotain postista (arvaus)* 'Said something about mail (a guess)' (Item 10: Z4:4, A)
- *Arvasin. Puhuttiin nukeista ja leikkimisestä* 'I made a guess. They talked about dolls and playing' (Item 11: X2:5, B→)
- *Arvaus, ainakin puhui taidenäyttelyistä.* 'A guess, at least she talked about art exhibitions' (Item 25: B12:8, B)
- *Arvaus, tartuin siihen kahviin* 'A guess, I grabbed that coffee' (Item 25: F2:3, A)
- Arvaus. Prix-sana esiintyi 'A guess. The word prix appeared' (Item 26: B10:6)

Type 4 guessing: Partial comprehension[76]

Despite having partially understood the spoken text, some test-takers still feel they have to rely on guessing in their selection of an option. The test-takers have actually understood bits and pieces of information given in the text, but have not felt confident enough to be able to directly discard the distractors and pick the key. In fact, another typical item constructing strategy is to have one of the distractors attract test-takers that have understood some, but not all of the essential information or the message in the text. Among the responses given to the seventeen items, we have 17 cases of evidenced partial comprehension.

[75] See ch. 9.3.1 for more details on the word-bound responses.

[76] See chapter 9.3.2 for more details on the partial comprehension-responses

Slightly more than half of these test-takers have arrived at the key – their success probably depending on exactly what part of the text they have understood: a part of the central message or some secondary piece of information. The following responses show partial comprehension:

- *Puhuttiin kauppojen sulkemisesta. (Arvaus)* 'They talked about closing down shops. (A guess)' (Item 1: O4:3, B→A)
- *Arvaus, halusi olla rauhassa* ' A guess, wanted to be alone' (Item 4: S8:6)
- *Arvasin, sillä nainen puhui vain siitä kuinka alun jälkeen nyt on mennyt paremmin* 'I made a guess, since the woman just talked about how after the start now everything has been better' (Item 6: O9:6)
- *Arvaus, taidegalleria on tehnyt jotain uutta näyttelyssään?* 'A guess, the art gallery had done something new with their exhibition?' (Item 25: S12:7, B)

Type 5 guessing: Unsure test-takers

There are test-takers who, perhaps due to personality factors, have simply not felt certain enough about their selection of an option and have therefore called their strategy "guessing" – they seem to give proof of self-monitoring and self-evaluation (See Second Language Listening Comprehension Inventory by Young 1997 presented here in chapter 1.6.3). Sometimes they have actually understood most of the text but for some unclear details. Their understanding may in some cases be sufficient for selecting the key. In fact, 57 % of these 21 participants have chosen the key, which may be taken as a proof of that situation. Both choices of a distractor and correct choices are exemplified by the following:

- *Arvasin, kun en ollut varma*' I guessed since I wasn't sure' (Item 3: Z7:4, A→)
- *En ole varma. Puoliarvaus.* 'I'm not sure. A half-guess' (Item 25: R11:8, B→)
- *Halvt gissade* 'I made a half-guess' (Item 27: H9:8)
- *Aika varma veikkaus* 'A pretty certain guess' (Item 28: S7:6)
- *Puoliarvaus...* ' A half-guess' (Item 28: L3:8, C)

Type 6 guessing: Problems with the options

Cases where the validity of some of the items is threatened by construct-irrelevant variance (difficulty) occur when the test-takers have had to rely on guessing due to problems in understanding either one keyword or the meaning of the written stem (question), or (some of) the options. We have a total of 27 such cases. This type of guessing is used by test-takers who may not be able to prove their listening comprehension because of an insufficient mastery of the vocabulary (or deficient reading comprehension ability). The essential question is whether the targeted listening comprehension ability construct covers this written vocabulary knowledge or whether the needed abilities are construct-irrelevant. In a valid and reliable test, construct-irrelevant abilities should affect the test scores as little as possible.

Here 41 % of the test-takers using this type of guessing may have missed the key because of an opaque question or unclear options. The same problem is evoked in the study by Yi'an (1998). There may naturally be other causes added to this factor – like listening comprehension deficiencies – that affect the process of response selection. Interestingly, the largest number of this type of cases occur for item 8, an item that otherwise has proven to function well (being on a convenient facility level, and discriminating sufficiently well). The judgements on the quality of item 8 therefore may have to be revised. In the following examples from several items, the construct-irrelevant variance may be judged to be a threat to the validity of the item:

- *Arvaus (Phénomène?)* 'A guess (Phénomène?)' (1: B6:7)
- *Ymmärsin kyllä mitä nauhalla sanottiin, mutten ymmärrä vastausvaihtoehdoista B & C kohtia, B vahvempi arvaus* I did understand what was said on the tape, but I don't understand options B and C, B is a stronger guess (1: U7:7)
- *Arvaus? Pientä tietoa. En ymmärtänyt kunnolla vaihtoehtoja.* (5: J3:3, A*)* 'A guess? Little knowledge. I didn't understand the options properly
- *Arvasin, koska en tiennyt mitä muut vaihtoehdot tarkoittavat.* 'I made a guess because I didn't know what the other options mean.' (Item 6: J8:6, A)
- *Mikä on cambriolages? Täysi arvaus* 'What is cambriolages? A complete guess' (8: S1:2)
- *Arvaus. En ymmärrä kysymystä.* 'A guess. I don't understand the question' (8: B7:5, A)
- *Arvaus (Inconvenient?)* 'A guess (Inconvenient?) (Item 9: B6:7, B)
- *Arvasin…sans frapper…? Ainoa joka tuntu oikeelta* 'I made a guess...sans frapper...? the only option that seemed right' (9: N9:8, A→C)
- *C:stä en oo ihan varma suomeks, joten arvon vähän* 'I'm not sure about C in Finnish, so I guess a bit' (10: U1:3, B→)
- *Ei oo varma, en ymmärtäny C kohtaa, eli hyvä veikkaus vaan.*'I'm not sure, I didn't understand option C, so it is just a good guess' (25: B5:6:B)
- *En muistanut kaikkien sanojen merkitystä, arvaus* 'I didn't remember the meaning of all the words, a guess'(27: B3:4 :A)
- *En ymmärtänyt a:ta, ja arvoin sitten b:n ja c:n väliltä* 'I didn't understand A, so I guessed between B and C'(30: V5:4)

Type 7 guessing: Ruling-out options

Type 7 of guessing is frequent: as many as 70 test-takers have stated that they have been able to eliminate one distractor (or through a misunderstanding perhaps the key), but have made a guess between the other two options. Some test-takers have explicitly stated that they have used the ruling-out strategy[77]. This combination of ruling-out and guessing is interesting as this phenomenon is probably only found when the MC item format is used. In the study conducted on a MC test of reading comprehension, Rupp (2006: 464) found evidence of this type of a combination of the two strategies. In his study, guessing was seen as the last resort and was only exerted upon those choices that were left after the knowledge-based or logic-based elimination of a few had already taken place.

However, of the present test-takers, less than one third (30 %) have been successful. Examples of this strategy, where the test-takers have either made a guess between two options, or applied the ruling-out strategy and made a guess, emerge from the following:

- *Ei ainakaan voi olla A joten B tai C. C on arvaus...'* At least it can't be A so either B or C. C is a guess…' (Item 1: B12:8, C)

[77] All the cases where the introspective responses indicate that the strategy of elimination has been used are treated below, in chapter 9.3.6.

- *Vain kaksi tyttöä → b ei ole oikein. Puhuttiin valokuvista, arvaus* 'Only two girls → b is not correct. Talk about photos, a guess' (Item 7: U8:7, C)
- *En ole varma onko A vai C oikea, siis arvasin* 'I'm not sure if A or C is correct, so I guessed' (Item 9: T3:4)
- *Arvasin, koska vaihtoehdot hyvin lähellä toisiaan* 'I made a guess since the options are very close to one another'(Item 26: J8:6)
- *C ei ollut oikein, arvaus An ja Bn välillä* 'C was not correct, a guess between A and B'(Item 29: B3:4, A)
- *En ymmärtänyt a:ta, ja arvoin sitten b:n ja c:n väliltä* 'I didn't understand a, so I made a guess between b and c'(Item 30: V5:4)

Type 8 guessing: Influence of the second listening[78]

The test procedure included listening to the text twice, with a response required after the first listening, but with the chance to change the selection of the option after the second time. As a result, some test-takers admitted to guessing after the first listening, but say that the second chance to listen made them more certain about what option to select. The second listening probably completed the lacking comprehension of the spoken text for these 21 test-takers although they did not necessarily select the correct option even after the second chance. However, as large a proportion as 62 % (13 test-takers) did arrive at the key. Successful and unsuccessful choices are exemplified in the following responses:

- *Ekalla kerralla arvasin puoliksi, mutta toisella kerralla kuulin vastauksen. En tiedä mitä C tarkoittaa* 'At the first listening I made half a guess, but at the second I heard the answer. I don't know what C means' (Item 1: F5:9)
- *Ekaksi arvasin, luulen että puhui kalakaupasta* 'At first I made a guess, I think she talked about a fishmonger's' (Item 10: K6:5, A→)
- *Oli yksinäinen, mutta piti silti lapsuudesta, ensimmäisellä kerralla ehdin vain arvata* 'She was lonely, but liked her childhood, at the first listening I only had time to guess' (Item 11: T5:4, A→B)
- *Ekalla kerralla vain arvasin. Toisella kerralla muutin vastausta, koska se sopi paremmin.* 'At the first listening I just guessed. At the second I changed the answer, since it fitted in better' (Item 25: E1:2, C→B)
- *Eka kerta meni ohi, en oikein tiedä, arvaus* 'The first time I missed it, I don't really know, a guess' (Item 28: T1:1, B→)
- *Ensimmäisellä kierroksella arvasin, mutta päätin vaihtaa arvaukseni* 'At the first time I just made a guess, but I decided to change my guess' (Item 29: Z2:3)

Type 9 guessing: Unclear reasons

There are cases where the test-takers are unsure of why they have selected a particular option, and say that they may have guessed. These types of responses are understandable, since it is not an easy task to reflect on one's own test-taking or comprehension process to know why a particular option is selected. To explain it as guessing may feel like an easy way out, especially as the strategy of guessing was explicitly mentioned in the written test instructions. These ten uncertain test-takers, of whom six have arrived at the key, probably have

[78] The factor of the second listening is also treated below in chapter 9.3.7.

some other reason than guessing for selecting an option, a reason that is or is not available for reflection:

- *Ei nyt ihan arvaus, mutten osaa perustellakaan...* 'Not quite a guess, but I don't know how to justify it…' (Item 1: B5:6)
- *Arvasin, tai en muista millä perusteella päädyin vastaukseen* 'I made a guess, or I don't remember on what basis I arrived at this response' (Item 3: P5:3, B)
- *En osaa sanoa, arvasin tai päättelin* 'I can't say, I made a guess or I inferred' (Item 3: Q2:5, B→)
- *Puoliksi arvaus, ei kunnon perusteluita* 'Half a guess, no good explanations' (Item 9: B11:8)
- *Voiko tällaiseen muutakin kuin arvata? ☺* 'Is it possible to do anything else than guess to one of these? ☺' (Item 27: Z12:6)
- *En tiedä, ehkäpä arvasin* 'I don't know, maybe I made a guess' (Item 28: Q2:5)

Type 10 guessing: None of the options is good

Three test-takers feel that none of the options matches the spoken text and have therefore made guesses. Can we conclude that they have missed something essential in the text contents? If the options are just and fair, the understanding of the text contents should give away the correct option. Two of these test-takers who have made a guess because of a "non-match" situation have chosen the key option:

- *Jag tyckte inte att de sa något om någon av dehär, gissning* 'I don't think they said anything about anyone of these; a guess' (Item 4: H7:7)
- *Mikään ei tuntunut sopivan, arvaus* 'None of these seemed to be right; a guess' (Item 6: S9:6, A→)
- *Arvaus, ei mielestäni sanottu mitään vaihtoehdoista* 'A guess, I don't think anything was said about the options (Item 7: Z8:5, A→C)

The guessing continuum

Judging by the different types of responses including the element of guessing that are exemplified above, the strategy of guessing is heterogeneous. What does the hypothetical continuum from a wild, random guess to a good, informed guess look like? In Figure 20 different types of guessing described above are included. From a cognitive processing perspective, a logical relationship between these types of guessing strategies and understanding of the text can be noticed. Random guessing is more likely to occur at the "no understanding"- end, whereas partial comprehension implies informed guessing. The feeling of uncertainty can probably be experienced through all levels of guessing as a function of the test-takers' personality and of the degree of comprehension of the text and, importantly, also of the question and the options.

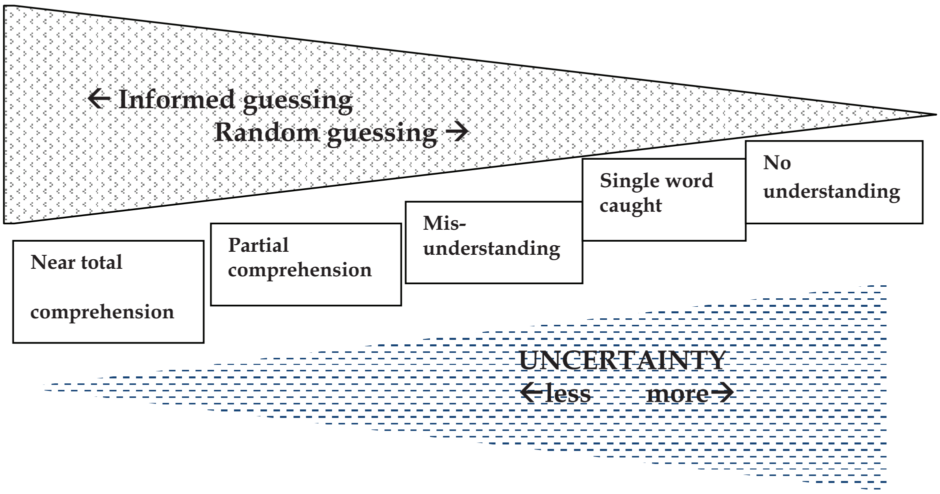


FIGURE 20 Guessing continuum

If we consider the relative success of the test-takers having applied these different types of guessing, the continuum is less linear, however (see Figure 21).

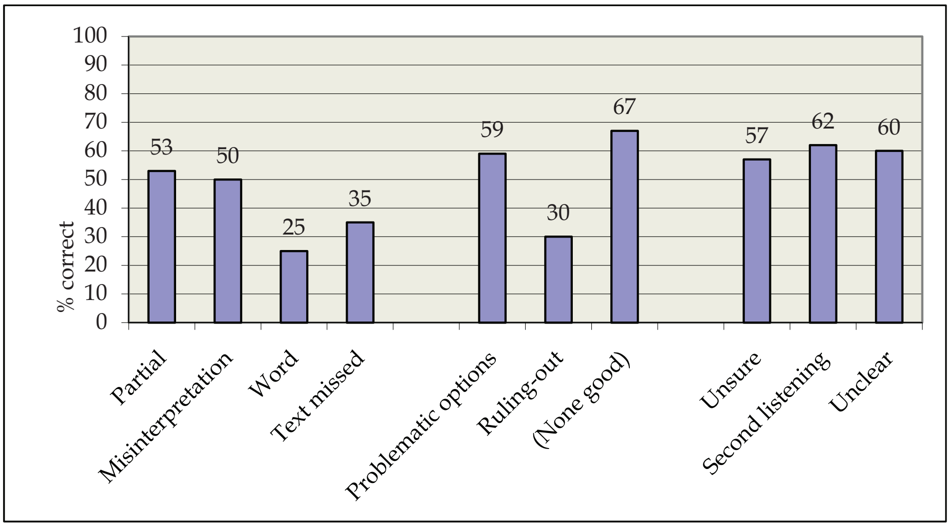


FIGURE 21 Type of guess and proportion of success

Considering the types of guesses that refer to some pieces of the spoken text (types 1, 2, 3 and 4), it is interesting to note that partial comprehension and misinterpretation have nearly equal success levels – close to 50 % - whereas guessing based on single words is definitely not a successful strategy, with a success rate of only 25 %. The proportion of correct guesses where the text has been missed reaches 35 %. We would perhaps expect a larger proportion of correct guesses among the partial comprehension-type. However, the situation of 53% correct responses may indicate that the partial comprehension has not consisted only of parts of the necessary information, but also of secondary information. Thus, in some cases, the bases for guessing have not been sufficient to give a correct answer. In fact, among all the responses indicating partial comprehension (N: 408) in the study as a whole, the proportion of correct choices of options is comparable, at 47.5 %. Where the focus is on the options (in guessing types 6, 7 and 10), even if the test-takers have experienced problems with the items, nearly 60 % of the test-takers have still arrived at a selection of the key.[79] Slightly surprisingly, the combination of elimination and guessing has not been a secure way to arrive at the key; only 30 % have succeeded. This seems to be a sign of the fact that these test-takers have based their elimination and guessing on a rather restricted comprehension of the spoken text.

The type of guessing that is due to the test-takers' feeling of uncertainty (type 5) shows that they have, in fact, probably understood more than they believe. It can also be a question of an influence of personality factors (see chapter 1.7). The types where the second listening has had an impact (type 8) and where guessing has been stated as a reason when no other justification has been clear (type 9), have, rather as expected, been reported with relatively high proportions of correct responses – around 60 %.

In the following, two more variables in the study of these responses and the differences related to these are considered: the separate items as well as the test-takers' success on the seventeen listening items.

Guessing at different items

In Figure 22 is presented the total number of guesses for the different items, and the combination of this strategy with a correct or an erroneous option selection.

[79] As there are only three test-takers who have experienced that none of the options is good (guessing type 7), the high percentage of success (67 %) is misleading and cannot be used as a basis of generalisations, as it concerns only two test-takers

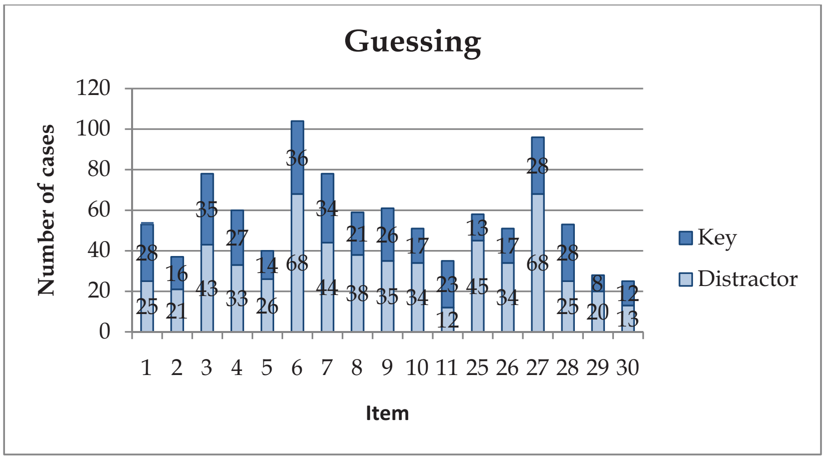


FIGURE 22 Total number of guessing and success per item

As far as the strategy of guessing related to separate items is concerned, the largest number of responses containing the element of guessing is given with item 6: in as many as 104 cases, the test-takers have indicated that they have guessed. The great majority of these guesses – 94 cases – are so called pure guesses with no further specifications given by the test-taker, which gives very little to rely on for a researcher who can only speculate on the reasons for or the ways of using the strategy. These guesses have to be considered in the larger context of all reactions with the entire selection of introspective responses to this item. In this comparison (see Figure 26 in chapter 9.5), where the categories are determined as described in the hierarchy (see chapter 9.1), 43.5% of all the test-takers say that they have chosen a particular option is by guessing the answer, without giving a further explanation. The low success rate (35 %) reached by these guesses suggests that in many cases they have, probably, been random.

Guessing seems to be used frequently as a response strategy in situations where the options are opaque, that is unclear, not unambiguously wrong or correct, or if they contain unfamiliar expressions. Item 6 is testing, among other things, the ability to compare negated – i.e. relatively more demanding - written statements with the text content, and the attitudes of the speaker towards a situation, expressed mainly verbally, but also by the tone of voice. To the complications adds the fact that the necessary information for item 6 comes before the necessary information for item 5.

Item 6 is a difficult item, with the item measure 55.93. There is probably something unclear with the options for item 6, since there are very few text-related responses and a vast majority of instances of the use of the strategy of guessing. Generally strong test-takers may have landed on a distractor, and weaker test-takers may have made a lucky guess: the reliability of the test scores for this item can be questioned.

Examples of introspective responses with the element of guessing include:

- *Arvasin, en kuullut* ‘I made a guess, I couldn’t hear’ (Z8:5, A)

- *Arvasin, koska en tiennyt mitä muut vaihtoehdot tarkoittavat* 'I made a guess, since I didn't know what the other options mean' (J8:6, A)
- *Mikään ei tuntunut sopivan, arvaus* 'Nothing seemed to suit, a guess' (S9:6)

Item 27 (item measure 58.30, S.E. 1.56, being the next most difficult item) follows in frequency with 96 cases, and an even lower success rate: 29 %. Whereas the majority of these guesses are indicated as pure guesses, with no further explanations, there are also cases where the test-taker has relied on guessing due to having missed the text, or misinterpreted it. Some have unsuccessfully combined guessing and ruling-out. These reasons probably also lie behind the other unexplained guesses and the generally low success rates for this item. Examples:

- *Jotain puhuttiin kirjeen lähettämisestä. Arvaus.* 'They said something about sending a letter. A guess.' (O13:7, C)
- *Miehen viimeinen lause oli jotain "passe", joten arvasin sen perusteella* 'The man's last sentence was "passe" something, so I guessed on the basis of that' (O9:6, B→C)
- *Va? Främst en gissning* 'What? Above all a guess'(H10:8)
- *Veikkaus B:n ja C:n välillä. A tuntui väärältä.* 'The guess is between B and C. A seemed wrong' (N2:4, C)

Items 3 and 7 both have a total of 78 cases of guessing, most of them pure guesses. The strategy of guessing has been more successful with these two items than with the two previous, with a success rate of around 45 %.

- *Han berättar att han jobbar ute. Gissade.* 'He tells that he works outdoors. A guess. (Item 3: D2:4)
- *Arvaus, mutta jotain tietoa sinnepäin* 'A guess, but something like that' (Item 3: J3:3, B)
- *Suljin pois vaihtoehtoja ja arvasin kahdesta* 'I ruled out options and guessed out of two' (Item 3: Å8:7, B)
- *Aika arvaus ku en täysin ymmärrä ku A-kohdan ja se ei vissiin oo se* 'Quite a guess as I only understand option A properly, and I guess it's not that one' (Item 3: B5:6)
- *Ajattelin, että vastaus ei olekaan C joten veikkasin A:ta* 'I thought the response is not C after all so I made a guess on A' (Item 7: Z2:3, C→)
- *Ensimmäisellä kerralla en ymmärtänyt mitään, toisella sain kiinni ideasta. Vastaukset sen sijaan ovat tod.näk. päin mäntyä. Arvasin ne kuulemani perusteella.* 'The first time I didn't understand anything, at the second I got the idea. The answers are likely to be all wrong, though. I guessed on the basis of what I heard.' (Item 7: P7:4, C)
- *Har två döttrar. Gissning.* 'Has two daughters. A guess.' (Item 7: H6:6)

On the other hand, the lowest number of guesses is for item 29 (28 cases in all), followed by item 30 (35 cases). Apart from pure guesses, for both these items the guessing strategy is combined with elimination in six cases.

- *Ensimmäisellä kierroksella arvasin, mutta päätin vaihtaa arvaukseni* 'At the first time I just made a guess, but I decided to change my guess' (Item 29: Z2:3, B→C)
- *Ensimmäinen vastaus ei sopinutkaan. Aika arvaus tämäkin.* 'The first answer wasn't good after all. This one is also rather a guess' (Item 30: J2:2, B→A)

Item 11 also has only 35 cases of guessing followed by 37 cases for item 2 and 40 cases for item 5. All these items have the common characteristics of being relatively easy items (item measures: 36.5, 47.4 and 49.2). Apart from item 11, a very easy item, they also discriminate well.

- *Puhuttiin rahasta…arvaus* 'They talked about money... a guess' (Item 2: J11:8, A)

- *Ljud störde henne, det lät som någon av dehär, gissning* 'Noises bothered her, it sounded like one of these, a guess' (Item 5: H7:7, B→C)
- *Eka oli arvaus, luulen että se silti oli A* 'The first one was a guess, I think it was A after all' (Item 11: J7:5, B→)

Two hypotheses can be made on the pattern of the relationship between the uses of different types of guessing with the item measure:

1) More frequent cases of guessing at more difficult items should be expected, i.e. a positive correlation between the number of cases of guessing and the item measure.
2) There is probably a higher proportion of correct guesses with easier items, i.e. a negative correlation between the percentage of correct guesses and the item measure.

A Pearson correlation gives at hand that the first hypothesis seems to be verified: there is a significant correlation at the 0.05 level between the item measure and the number of cases of guessing (See Table 27 in Appendix 4). There is an even stronger correlation along the second hypothesis: a significant correlation at the 0.01 level between the item measure and the proportion of correct responses.

Conclusions on the strategy of guessing

Even if the results of the use of guessing in this study only represent a sample of cases, they reflect the complicated outcomes of the test-takers' processes used in a test-taking situation. The critique against the MC format in general, associated with the risk of a lack of reliability and thereby possibly a lack of validity, appears to be legitimate. On the other hand, we can see that there is more than meets the eye when it comes to guessing on MC items. This was also found in the study on a MC test of reading comprehension where Rupp (2006: 464) explains that guessing could not generally be characterized as an uninformed process whereby a random selection is made among the possible choices, but rather as conditional informed guessing. Guessing is only rarely random guessing, and it is not automatically a bad strategy, always associated with the MC item format. The strategy of guessing often implies inferring, reasoning and elimination, ingredients of many communicative language use situations (Cf. Bachman & Palmer 1996; Linn & Miller 2005).

The conclusion that can be drawn from the present analysis of guesses relate to the different dimensions of the processes and strategies activated in the situation of a test of listening comprehension involving the spoken text, the questions (items, stem and options) and the individual test-taker. If the spoken text is difficult to understand, if the test-taker has only grasped fragments of it, the tendency to guess is naturally higher, since the basis on which the question has to be answered is less stable. The contrary has to be true also: if more is understood, if the text is experienced to be easier, the questions are more easily answered correctly.

However, as is evidenced in the introspective responses, the questions and the options may add to the possible confusion. If the question is difficult to understand, if the task demanded is too difficult for the targeted level or if the written language is too complicated, thus causing the task to become opaque, the test-taker has to rely on guessing, possibly combined with elimination, to solve the task. Similarly, Rupp (2006) found that for reading, the main deciding factor for choosing a strategy was the perceived difficulty of the text OR the questions.

We can assume that to the decisive factors of the text and the task adds the characteristics of the test-taker, and the level of certainty or uncertainty she experiences when faced with a specific task. Different test-takers seem to handle challenges in different ways. A test-taker with fairly low comprehension ability but who is confident by nature may think he or she knows how to reach the answer, even if his or her interpretation of the text would be erroneous. Some test-takers are risk-takers who don't mind hazarding on the basis of whatever textual or test-wise clues they have at their disposal. The test-taker may not experience that he or she has guessed, or at least not admit to it. On the other hand, a less confident test-taker, even if his or her comprehension ability is high, and even if he or she has understood most of the spoken text, may feel the need to guess if the text is not completely understood or if he or she is uncertain as to the meaning of some details in the text, the question or the options[80].

It can be hypothesised that, in general, the situation where the stronger the test-taker, the less frequent are the guesses, holds true. This goes along the lines of reasoning where a correct response is arrived at by means of understanding the text. If we take all different introspective responses given for this current test into account, the highest proportion of guesses ("pure guesses") among all the different responses given on a particular level is found among the test-takers with the lowest person measures. The proportion is decreasing towards the test-takers with higher person measures. This is also shown as a correlation between the number of pure guesses[81] (See Table 25 in Appendix 4) and the person measure. The higher the person measure, the lower is the number of guesses among the individual introspective responses.

The implications for test construction that can be concluded on the basis of the results obtained through this study follow the ideas of Bachman and Palmer (1996: 205), who recommend that provisions for eliminating or reducing the potential causes of random guessing be included in the stage of test design. These are: providing ample time for the majority of the test-takers to complete all the tasks in a test, matching the difficulty of the items with the ability levels of the test takers and encouraging test-takers to make informed guesses on the basis of partial knowledge. To these can be added, on the basis of the present study, the importance of creating transparent questions (stems and options) so

[80] This is further discussed chapter 9.3.7 on the metacognitive responses.

[81] "Pure guesses" represent the introspective responses that include only the explicit mentioning of a guess, without further explanations. In this current chapter focusing on the strategy of guessing, to these add all the other responses that include the notion of guessing but with an added hint as to the basis or the reason for guessing.

that they are understood and interpreted correctly by the test-takers, since they clearly influence the processing of the text.

9.3.6 Strategy-based responses: Elimination This chapter will focus on the frequent cases of more or less explicit indications of uses of the strategy of ruling-out or elimination. The questions that I want to address are:

- In what ways do the test-takers apply the strategy of elimination?
    - What subtypes of elimination or what ways of expressing the use of this strategy are there?
    - How frequent are these different subtypes in relation to the total amount of cases of this strategy?
- Is there a relationship between the test-takers' success on the individual items and the use of a particular subtype of elimination process?
- Is there a relationship between the level of difficulty of an item and the use of the elimination strategy in general or a particular subtype of elimination?

The answers to these questions will enlighten the nature of the process of elimination from the point of view of the justifiability of the use of the process as it can be taken as a part of the listening comprehension construct. The hypothesis I have is that the introspective responses will prove that elimination is not in itself automatically a "bad strategy" incompatible with the construct of listening comprehension in a language use situation. Consequently, the criticism against the MC test format from the point of view of the test-takers' use of strategies and processes that are not part of the construct of "natural" language use will not be entirely justified. However, as analyses on the process of guessing have indicated, the quality of the test and the individual items are essential: opaque (unclear) items are likely to promote the use of "bad strategies" – e.g. random guessing or elimination based on superficial reasons - where the result of a test-taker is not dependent on the efficient use of construct-relevant processes, but construct-irrelevant ones.

I will start by addressing the two first questions, concerning the nature and the frequency of the subtypes of elimination strategies that have been employed in this particular test situation.

Subtypes of elimination in the test of listening comprehension

Under the category of elimination, I have grouped several different subtypes of cases that give indications of more or less conscious use of the ruling-out strategy. These subtypes seem to be on different cognitive levels. I am well aware of the fact that for this sub categorisation, there are at least three somewhat problematic issues to be taken into account. The first one is the fact that the test-takers choose to write down what they want, meaning that they are not required to or necessarily able (partly because of time restrictions) to write complete responses, but they express their focus or what they consider pertinent in the particular situation. Secondly, related to this fact, the researcher has to rely on subjective interpretations of the test-takers' responses. There is thus often more thorough processes and reflections behind a response than the researcher is able to deduce. The processes and strategies with the focus set on elimination are parts of the complicated test-taking process (see Figure 14 in Appendix 3), and reflect different stages of this process, from the comprehension and the interpretation of the text, to the comprehension of the question and options, to the selection task. Thirdly, some responses include elements that would suggest a placement in two or even more subcategories. However, the choice is made to label each response by only one subcategory in this context, in order to be able to compare the proportions between test-takers with different persom measures, for example.

There are altogether 583 instances of the ruling-out strategy among a total number of 3706 introspective responses, giving at hand that as high a proportion as 16 % (15.7 %) of all of the responses given to the seventeen MC items contains the element of elimination[82]. In Table 28 are given the 13 subtypes of cases of elimination in the order of their frequency, the number of cases of each subtype as well as their proportion among the total number of cases:

TABLE 28 Subtypes of elimination, their number and proportion among all cases of elimination

	Subtype of elimination	**Number of cases**	**Proportion among all cases of elimination (N: 583)**
I	2 options are wrong	134	23 %
II	Best option	105	18 %
III	Elaborated elimination	63	11 %
IV	Quotation from text	51	8.5 %
V	One option eliminated + guess	39	6.5 %
VI	Explicit mention of strategy	35	6 %
VII	Comprehension problems	32	5.5 %
VIII	Nothing said about the other options	29	5 %
IX	Elimination on meta-level	28	4.5 %
X	One option eliminated	28	4.5 %
XI	Only possible option	24	4 %
XII	Change explained	11	2 %
XIII	No good option	4	0.5 %
	Total	583	100 %

The clearly most frequent subtype of elimination is the case where two options have been considered wrong (subtype I, with 23 % of all cases of elimination), followed by the subtype where one option is considered the best one (subtype II, at 18 %). These responses are rather vague, however, as they do not reveal on what basis the judgments have been made. The subtypes that are, on the contrary, more explanatory are the 'elaborated responses' (subtype III) and the 'qu-

[82] In the original categorisation, some of these 583 responses are placed within another category than option-focused responses, since they contain text-related elements or metacognitive reflections, which are placed higher in the hierarchy for categorising the introspective responses. This leaves a proportion of 14 % of option-focused responses in the original categorisation. (See Figure 13, chapter 9.1).

otations' (subtype IV), where there is a more clear stance taken towards each of the options and their correctness in relation to the text. The cases of 'elaborated elimination' and 'quotations' reach proportions of around 10 % (11 % and 8.5 % respectively). Most of the other subtypes are represented by around 5 % of the cases.

I will now proceed with my subjective definition and description of each of these subtypes, giving examples of responses that the test-takers have written down for different items. I will also comment on the possible background and consequences to the employment of these subtypes of strategies.

Type I: 2 options are wrong

The most frequent subtype of response containing the element of elimination is the one where the test-takers indicate that they have selected an option since the two other ones are wrong. The question arises whether the test-takers have not considered the key option to be clearly correct in case if they have been forced to proceed through establishing that the two others are wrong[83]. On the other hand, it may be a case similar to the MC test of reading comprehension, described by Rupp (2006), at which the test-takers generally assumed that all response options had to be read, understood and eliminated before the correct option could be selected. Examples are found in:

- *Muut vaihtoehdot huonoja* 'The other options are bad' (Item 2: M2:5, B→)
- *Mielestäni A tai B ei ollut oikein, joten vastasin C.* 'I thought A or B wasn't correct, so I answered C' (Item 3: B11:8)
- *Koska muut ei tunnu oikeilta. (Karsii väärät)* 'Since the others don't seem correct. (Sorts out the wrong ones) (Item 4: V9:6)
- *De två andra alternativen passade inte* 'The two other options were not appropriate' (Item 8: H2:4, B)
- *Muut vaihtoehdot on tyhmiä* 'The other options are stupid' (Item 8: L3:8)
- *A ja B eivät ole loogisia, C kai menettelee.* 'A and B are not logic, I guess C is all right' (Item 30: Å14:8)

Type II: Best option

The next most frequent subtype of elimination (at 18 %) consists of the cases where the test-takers indicate that the option they have selected is simply the best one. They tend to use superlatives of adjectives like 'good', 'suitable', 'probable' implying that a comparison has been made between the options. Does this imply that they do not consider any of the options being a convincingly good one but that the selected one appears to be the best among the proposed ones, or is it just a way of saying that they do not exactly know the reason for selecting a particular option? In both cases, it seems that the test-takers have considered the three options, compared them and arrived at a conclusion on which one to select (Cf. Rupp 2006). This subtype is exemplified by:

- *C on vaihtoehdoista sopivin* 'C is the most suitable of the options' (Item 2: X3:5)
- *Tuntuu todennäköisimmältä vaihtoehdolta* 'Seems like the most probable option' (Item 6: O8:6)

[83] The proportion of choices of the key option among this subtype of responses is given and discussed further below.

- *Tuntui oikeimmalta* 'Seemed to be the most correct one' (Item 9: Q2:5)
- *Näistä paras vaihtoehto mielestäni* 'The best option of these ones in my opinion' (Item 9: S7:6)
- *A vaikutti parhaimmalta* 'A seemed to be the best one' (Item 26: Z2:3, A)
- *Kaikkein mahdollisin vaihtoehto mielestäni* 'The most possible option in my opinion' (Item 28: V8:5)

Type III: Elaborated elimination

I have considered this subtype of responses consisting of the most "high-level" elimination, where the test-takers consciously seem to take into account each of the different options while considering the text contents. This seems to be a justified way of using the strategy, and a test-taking process that is actually intended by the test constructor to be employed at this specific test format. Hypothetically, this happens as follows. After having got a focus for the task by reading through the MC question, the test-taker listens to the text and forms an inner representation of the contents in working memory. She then compares this representation with each of the options, establishing the correctness or falseness of each of them. This description can be compared with the information processing model by Jamieson et al. (2000; here referred to in chapter 2.4.2). Obviously, the more the test-taker understands of the text, and the more exact an interpretation he or she has formed of it, the clearer the message will be, and the easier it is to compare the message with the representations that are made of the contents of the options. If the interpretation of the text is correct, the key option should be selected as an obvious answer to the item in question. Here this subtype of elimination, representing 11 % of all cases, is exemplified:

- *Puhuttiin kaupoista, enkä ollut varma mitä B:ssä tarkoitetaan, mutta asiayhteyksistä päättelin että vastaus on B. Plus että en kuullut puhuttavan pienten kauppojen loppumisesta* 'They talked about shops, and I wasn't sure what they meant in B, but from the context I could infer that the response is B. Plus I didn't hear anything about the closing-down of small shops.' (Item 1: V9:6, A→)
- *Ovi oli auki – ei b, ei varastettu – ei a, c kuulostaa yhteneväitä tekstin kanssa* ' The door was open – not B, not stolen – not A, C seems to be in accordance with the text' (Item 8: E11:9)
- *Olisin valinnut C:n, mutta näkihän se joskus äitiään. Ihmisiä oli koko ajan ympärilllä, eli B ei sitten ollut →päädyin A:han* 'I would have selected C, but she did see her mother sometimes. There were people around all the time, so it wasn't B then → I ended up with A' (Item 11: N10:9)
- *Se on kissalle →A ei käy. Hän ottaisi filettä →C ei käy* 'It is for the cat → A is not good. She would like a filet → C is not good' (Item 29: Å2:4)

The relative success of this particular subtype of elimination will be treated further below. The hypothesis is that this should be a successful strategy in general, in most cases leading to the rejection of the distractors, and the selection of the key.

Type IV: Quotation from the text

Quite a few test-takers (8.5 %) justify their ruling-out of particular options by referring to words, phrases or idea units that they have grasped in the text. (The more complete interpretations where each of the options is considered in the light of the text are placed in the category of elaborated elimination). The success of the test-takers who have given these particular responses varies according to the amount of text they have understood and the correctness of their interpretation of that text, as well as according to their ability to relate their interpretation to the contents of each of the options. Examples of this subtype of elimination include:

- *Tekstissä puhuttiin kaupoista, joten muut vaihtoehdot eivät tulleet kyseeseen* 'They talked about shops in the text, so the other options could not be possible' (Item 1: J8:6, A)
- *A ei ainakaan. Toisella kerralla tajusin kuuntelusta enemmän. Jotain laihtumisesta.* 'At least not A. At the second listening I understood more about the text. Something about losing weight' (Item 5: U4:5, C→)
- *Muita vaihtoehtoja ei mainita.(Puhuttiin kaupasta)* 'They didn't mention the other options. (They talked about a shop)' (Item 10: Q1:4)
- *"Etkö tiennyt sen olevan pakoll.?" → B ja C ei käy* '"Didn't you know that it is compulsory? → B and C are not good' (Item 28: N9:8)

Type V: One option eliminated + guess

A proportion of 6.5 % of all test-takers using elimination say that they have ruled out one of the options, and have made a guess between the two remaining ones.

- *Ei ainakaan voi olla A, joten B tai C. C on arvaus…*'At least it can't be A, so B or C. C is a guess…' (1: B12:8,C)
- *Poissulkien + arvaamalla* 'By ruling-out + guessing' (7: J7:5, C)
- *Myös A tuntui hyvältä vaihtoehdolta, mutta päädyin B:hen.* 'A felt a good option also, but I ended up with B' (25: E2:2, B)
- *Arvasin, en osannut päättää B:n ja C:n välillä* 'I made a guess, I couldn't decide between B and C' (25: N6:6, B)
- *C ei ollut oikein, arvaus An ja Bn välillä* 'C was not correct, a guess between A and B' (29: B3:4, A)

Type VI: Explicit mention of strategy

Some test-takers (6 %) have been able to name the strategy they have used, apparently being conscious of both the existence of this strategy and their use of it in that particular situation – through self-monitoring. An interesting factor to investigate is the success of the test-takers who have explicitly mentioned this strategy: are these the most successful ones, or is the success completely random, or perhaps the conscious use depends on factors like test experience? Here are examples of this subtype:

- *Uteslutningsmetoden* 'Ruling-out method' (Item 1: H12:10)
- *Käytin poissulkemistaktiikkaa* 'I used the ruling-out tactics' (Item 4: E6:6, B)
- *Suljin pois vaihtoehdot yksitellen* 'I ruled out the options one by one' (Item 7: Å8:7)
- *Poissulkemismenetelmä* 'Ruling-out strategy' (Item 25: Z4:4)
- *Sulkemalla aluksi väärät vaihtoehdot pois* 'By first ruling-out the wrong ones' (Item 30: Y1:7, A)

VII: Comprehension problems[84]

As large a proportion as 5 % of the option-focused responses reflect the situation where the test-takers' problems of understanding the options influence their use of elimination. Sometimes this has lead to the rejection of the problematic option or options, whereas in other cases it has been the opposite: a difficult word in an option has been taken to indicate that it is the key. In an ideal case, if the text has been correctly interpreted, the key option should give itself away easily. However, if there are problems in understanding either the text or the questions (stem or options) this will not be a simple task. From the point of view of the quality and reliability of a test, in cases where the spoken text has not been understood, the choice is expected to be a distractor. However, the validity and the reliability of the test are at stake if the test-takers' comprehension of the written options determines their success on the test items. Various problems are reflected in the following examples:

- *Asunto oli kallis, enkä ymmärtänyt B-kohtaa...* 'The appartment was expensive, and I didn't understand option B' (Item 6: V5:4, C)
- *En kuullut mainittavan mitään A:tai C:hen liittyvää, joten otin B:n, jota en ymmärrä* 'I didn't hear mentioned anything related to A or C, so I took B, that I don't understand' (Item 6: N12:11)
- *Ainoa vaihtoehto jota en ymmärtänyt, ja muita ei mainittu tuolla tavoin* 'The only option that I didn't understand, and the others were not mentioned like that' (Item 8: S9:6, A)
- *A:skulle vara för enkelt om det var svaret, C:visste inte vad "prix" betydde.* 'A: would be too simple if it was the response, C: didn't know what "prix" meant' (Item 26: D2:4, B)

Type VIII: Nothing said about the other options

There are cases where the test-takers claim that they have selected an option because nothing else is said in the text about the other options. Here weaker students searching word-matching are most likely to arrive at a distractor. However, if the match is checked against the overall contents of the text, the probability of success is greater. This will be verified when comparing the success rates of the test-takers having employed this subtype of elimination. Examples of this subtype include:

- *Nauhalla ei mainittu a eikä b kohtaa ja päädyin c:hen. Kuulin hyvin tämän kohdan toisella kerralla* ' On the tape, options A and B were not mentioned and I ended up with C. I heard this passage well at the second listening' (Item 2: U7:7, B→)
- *Muita mielestäni ei mainittu* 'I don't think the other ones were mentioned' (Item 5: R7:5)
- *Ei väittänyt A:ta tai C:tä* 'He didn't claim A or C' (Item 5: K15:7)
- *Hän haluaa muuttaa takaisin kotiseudulle keski-Ranskaan. Muita ei mainittu.* 'She wants to move back to her home region in central France. The other ones weren't mentioned' (Item 5: Q5:8, B→C)
- *A, koska muita ei sanottu selvästi.* 'A, since the others were not clearly mentioned' (Item 7: Q3:7, C→A)

[84] The issue of problems understanding the options is discussed further in the following chapter on meta-cognitive responses, illustrated with examples.

Type IX: Elimination on the meta-level

Elimination on the meta-level implies that a selection is justified by grammatical, semantic or test-strategic considerations. Examples of these justifications are found in:

- *Muut vaihtoehdot ei ole lähellä puheenaihetta* 'The other options are not close to the subject of discussion' (Item 2: O8:6)
- *Kuulin ettei ollut A, C kuulosti parhaalta, koska niin summittainen* 'I heard it wasn't A. C sounded the best, as it was so approximate' (Item 2: N8:7)
- *Joko A tai B, mielestäni vastauksen ei pitäisi olla imperfektissä joten B* 'Either A or B, I don't think the response should be in imperfect tense therefore B' (Item 28: E6:6, B)
- *Filettä ei ole enää, vastaus A olisi epäkohtelias* 'There is no more filet, option A would be impolite' (Item 29: T7:4)
- *En nyt tiedä, tässähän käy A ja B, jos nyt oikein vaihtoehdot ymmärrän suomeksi* 'Well, I don't know, here both A and B would go, if I understand the options correctly in Finnish' (Item 29: V6:5)

Type X: One option eliminated

Sometimes one of the options has clearly stuck out as a distractor for the test-taker and this one option has been ruled-out immediately, as expressed by the test-takers in the following examples:

- *Ei A* 'Not A' (Item 5: J4:3)
- *Ei ainakaan B* 'At least not B' (Item 8: N12:11)
- *Mielestäni A tai C ...ei ainakaan B* 'In my opinion A or C... at least not B' (Item 10: Z3:3, A)
- *Mielestäni A on sittenkin ainut mikä kävisi miehen kommenttiin. C se ei ollut ainakaan...* ' I think A is after all the only one that would go with the man's comment. At least it wasn't C…' (Item 29: V7:5, B→A)
- *Ei ainakaan B joten jäi 2 valittavaa. Päädyin sittenkin C:hen.* 'At least not B so I was left with 2 to choose between. I ended up with C after all' (Item 30: V7:5, A→)

Type XI: Only possible option

In some cases the test-takers express the reason for selecting an option with their feeling of it being the only possible alternative that is appropriate, logic or suitable, as exemplified in:

- *Ainut järkevä vaihtoehto* 'The only sensible option' (Item 3: E6:6, A)
- *Ainoa, mikä mielestäni sopis* 'The only one that is appropriate in my opinion' (Item 27: J2:2, C)
- *A oli ainoa looginen jatko keskustelulle* 'A was the only logic continuation for the conversation' (Item 28: Å15:8)
- *Ainoa, joka käy* 'The only appropriate one' (Item 30: R3:3)
- *Ainoa oikean kuuloinen vaihtoehto*. 'The only option that sounds correct' (Item 30: Y2:7)

Type XII: Change explained

A few test-takers have explained the reasons behind changing their option selection after the second listening of the text. The following subtypes of responses are found:

- *Periaatteessa toi A sanottiin myös mutta ei pelkästään* ' In theory A was also said but not only' (Item 1: R12:8, A→)

- *Arvaus, päätin vaihtaa, koska A tuntui luonnollisemmalta* ʼA guess, I decided to change, as A seemed the more natural one' (Item 29: R1:2, B→A)
- *C ei sovi sittenkään, joten A paras vaihtoehto.* 'C is not good after all, so A is the best alternative' (Item 29: J3:3, C→A)
- *Hahaa! Nainen halusi filettä kissalleen, eikä sitä ollut...joten B on sopiva. Korjasin.* 'Hahaa! The woman wanted filet for her cat, and there wasn't any left…so B is the suitable one. I corrected my choice.' (Item 29: L2:7, A→)
- *Ehkä A sopii sittenkin paremmin, innostuneempi reaktio* 'Maybe A is better after all, a more enthusiastic reaction' (Item 30: U7:7, C→A)

Type XIII: No good option

A handful of test-takers say that none of the options seems good. They may have misunderstood the text contents or the options, and thus find it difficult to match their representation of the text with the options:

- *Tuntui ettei mikään käy, vaihtoehto tuntui parhaimmalta* 'I felt as if nothing was good, this option seemed the best one' (Item 3: S9:6, A→)
- *Mikään ei tunnu hyvältä...* 'Nothing seems good…' (Item 6: Å14:8)
- *Mielestäni mikään ei oikein sopinut, mutta ehkä C parhaiten kai...* 'I don't think anything was really appropriate, but maybe C is the best one I suppose…' (Item 27: Z14:7, C)

Subtypes of elimination and success

The next question concerns the subtypes of elimination that are related to the most or the least successful outcome: what strategy has most frequently lead to the selection of the key option versus a distractor?

TABLE 29 Subtypes of elimination and proportion of correct responses (in descending order)

Subtype of elimination	Number of cases	Proportion of correct responses
III: Elaborated elimination	63	81 %
X: One option eliminated	28	71 %
IX: Elimination on meta-level	28	68 %
VIII: Nothing said about the other options	29	65 %
XI: Only possible option	24	63 %
I: 2 options are wrong	134	61 %
VI: Explicit mention of strategy	35	60 %
IV: Quotation from text	51	53 %
XIII: No good option	4	50 %
II: Best option	105	49 %
XII: Change explained	11	45 %
VII: Comprehension problems	32	44 %
V: One option eliminated + guess	39	36 %
Total	583	
	Mean: 45	Mean: 53 %

As expected, 'elaborated elimination' turns out to be the most successful subtype at this test: above 80 % of these cases have lead to the selection of the key (see Table 29). This subtype is followed by 'one-option elimination' and 'meta-level elimination' at a success rate of approximately 70%. An interesting result here is the difference between the success rates of the cases where one option has been rejected (subtype X) versus the cases where this is combined with guessing (subtype V). In the latter case, the success rate is the lowest of all subtypes of elimination - only 36 % and close to the likelihood of success at random guessing when the text is missed (see Figure 21). This suggests that this subtype of elimination is inclined towards the random guessing process more than the informed elimination process. The fact that the mean percentage of correct responses for all the different subtypes of the strategy of elimination taken together is above 50 % suggests that the strategy implies in general something else than random guessing or ruling-out based on construct-irrelevant random factors (like the order or wording of the options, for example). This is also the interpretation that can be made based on the general contents of the responses. The fact that the subtype where the test-takers have indicated comprehension problems with the options is found towards the end of the success list seems expected: this further strengthens the importance of creating transparent items, to avoid construct-irrelevant variance.

Based on the analysis presented above, three subtypes of elimination that may be considered to be on a higher cognitive level (defined as conscious, well-targeted processes or strategies, based on a more or less correct representation of the oral text) are 1) elaborated elimination, 2) elimination on meta-level and 3) one option eliminated. These seem to be higher-level strategies by their definition and by the success level of these subtypes, reaching near 70 % and above. Added to this, they are associated with test-takers with higher person measures.

Subtypes of elimination and the test-takers' success

Considering the fact that the category of elimination consists of subtypes that vary considerably to their nature, a relationship between the individual test-taker's scores and the frequency of the general use of the strategy of elimination is not expected. This is verified by a correlation analysis (Pearson): no significant correlation between the number of cases of elimination[85] and the individual test-takers' person measure (See Table 25 in Appendix 4).

However, there may be a tendency of some test-takers towards selecting a certain subtype of elimination strategy rather than another subtype. The subtypes of elimination that show a tendency of growing in frequency towards test–takers with higher total scores or person measures are those of 'elaborated elimination', 'quotations', 'elimination on meta-level', 'explicit elimination' 'one-option elimination' and the 'only possible option', as well as the responses indicating comprehension problems. On the other hand, for test-takers with lower total scores or person measures, the two most frequent subtypes employed are the 'best option' and the 'two wrong'- subtypes. These subtypes of elimination do not reveal very much about the processes on which elimination has been based. This suggests either that these test-takers have simply chosen not to write down any more complete information, or that the process of com-

[85] Here the category of elimination is taken to be represented by the original cases of option-focused responses, slightly fewer than when considering all the cases for the study where elimination is evidenced.

paring the text contents to the contents of each of the options has failed at some stage. The processing of the text may have been complicated due to unfamiliar phonology, vocabulary or syntax; the test-takers may have not been able to keep the representation of the text long enough or hold a clear or complete enough representation in mind for the test-taker to be able to compare the representation with each of the options; or perhaps the stem and the options themselves have been unclear. What is more, for a weaker test-taker the test task itself may have taken up the most of the response time, not giving them a chance to think very deeply about the supplementary task, or a chance to write down what they are thinking.

Relationship between the individual item and the use of elimination

Another meaningful approach to the process of elimination is to look at how elimination is used with the seventeen individual MC items. The total number of cases of elimination varies to some degree between individual items: the total number ranges from 17 to 57 cases per item, the mean being 34. For items 5, 7, 8, 9, 25, 27, 29 and 30 the total number of responses exceeds the mean (see Figure 23). There is hardly any common feature for these items; they are very different by subtype and have for example their item measure and discrimination indexes on different levels.

The total number of cases of elimination is therefore not expected to correlate to any degree with the item measure, whereas some significant relationship between the proportion of correct responses and the item measure is more likely. The results show that as can be expected, the proportion of correct responses shows a significant negative correlation (Pearson) with the item measure (See Table 30 in Appendix 4). This implies that the easier the item, the more correct responses among the cases of elimination. This is obviously a somewhat circular result.

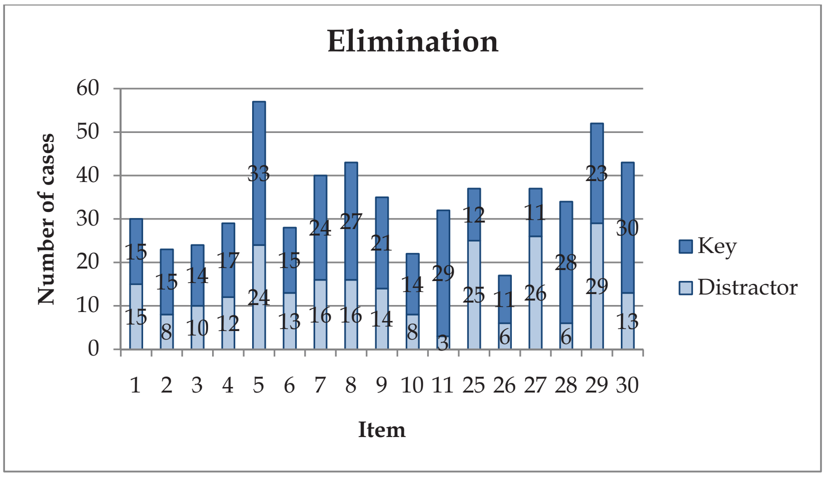


FIGURE 23 Total number of elimination and success at items 1-11 and 25-30

Conclusions on the strategy of elimination

There are some conclusions that can be drawn from this analysis into the strategy of elimination for a MC test of listening comprehension. First of all, the analysis seems to suggest that the claim put forth by for instance Ebel and Frisbie (1991) or Haladyna (2004) is justified: elimination can indeed be a high-level cognitive strategy, involving comparative judgments of the different options. However, depending on factors like the nature of the spoken text and the task or the individual test-taker's characteristics and experiences, the process can vary to a great extent. Some subtypes of elimination seem to be higher-level processes, leading to a more successful outcome, thus being associated with test-takers with higher person measures. In brief, elimination can be a useful and acceptable strategy of problem solving also in an assessment context.

However, if hardly anything is understood of the text, the basis for eliminating is poor and the strategy has to be based on other types of information. One or two of the options in an item may be implausible or the key option may seem better than the others for some reason, so the test-taker can base his elimination procedure and his selection on these factors. Elimination is described by Cohen (1998a) as representing such test-wiseness strategies, being compensating by nature. Sometimes the strategy of elimination is also combined with guessing.

When the comprehension problems are related to the question or the options, this creates a more complicated situation. The elimination and selection procedures become even less straightforward, and depending on the degree of comprehension of the text and some possible clues, the outcome will be more or less successful. Clearly, the further away we move from the focus on text comprehension, the greater the risk of construct-irrelevant variance.

9.3.7 Metacognitive introspective responses

Many test-takers have reflected on the items or on their test-taking when solving the seventeen MC items. The introspective responses categorised as metacognitive responses reveal further how the test situation has worked from the test-takers' point of view. The responses can be divided into seven (7) further subcategories according to what aspect in the process is focused. In some responses the test-taking situation is described more or less thoroughly (abbreviated to **SIT** in the tables). There are responses where the helpfulness of the second listening becomes clear (**2LIST**) whereas some responses show how, as a test-taking strategy, the test-takers have reasoned logically on the bases of various pieces of knowledge (**LOG**). There are responses that reflect problems with understanding the written options or the question (**OPT**) and others where the text has caused hesitations in the test-takers' interpretation process (**UNC**). Some responses, on the other hand, reflect the test-takers' confidence in solving a particular item (**CER**). Finally, some test-takers do not know how to explain their option selection (**??**). In the following, the seven types of metacognitive responses will be described, exemplified and discussed (See Table 31 in Appendix 4).

SIT: test-taking situation

The test-taking situation itself is referred to in quite a few responses. The test-takers reflect on their practical or emotional managing of the situation. This type of strategy is referred to as self-monitoring, self-evaluation or perhaps problem identification by Young (1997), or comprehension monitoring or evaluation by Goh (2002). However, while in their inventories these are described as being typically associated with higher-ability listeners, there are proportionally more of these responses reflecting features of the test-taking situation at the lower levels of scores in the current study, where there is also proportionally less correct responses. There are different reasons why weaker test-takers experience more anxiety with the task and succeed less frequently than stronger test-takers. Four main features are referred to in these responses: the test-takers' own deficiencies, the practical procedure of the test-taking, the basis for making the selection and the characteristics of the test.

Some test-takers reflect on the problems they are experiencing: trouble with the vocabulary, with finding the correct response, or with the ability to concentrate at the end of the test. This can be related to what is said about the nature of verbal protocols in general: more can be found out about the listening comprehension process when it does not flow comfortably because the processing becomes less automatic, but slows down to be more conscious (cf. Brown 1995, Yepes 2001, in chapter 3.1.3):

- *No niin se tuntu sit kuitenkin juttelevan. Ekal kerral arvasin ku iski paniikki* 'Well that's what she seemed to talk about after all. At the first listening I panicked and made a guess' (Item 1: L1:5, C→A)
- *Ihan pelkkiä arvauksia, liian huono sanavarasto niin ei ymmärrä puhetta* 'Just simple guesses, too bad a vocabulary to understand speech' (Item 3: P1:4, B)
- *En ymmärrä tekstiä täydellisesti, tiedän mistä puhutaan, MUTU-pohjalta ei ne vissiin ollu tunnettuja* 'I don't understand the text perfectly, but I know what they talk about, on the basis of a feeling I guess they weren't famous' (Item 25: J7:5,B→)
- *Arvaus, ymmärsin kuulemani mutten kuullut vastausta* 'A guess, I understood what I heard, but I didn't hear the response' (Item 26: Z1:2)
- *Ei ole paikalla, mutta miten se sanotaan?* 'Was not there, but how do you say it?'(Item 27: S1:2)
- *Toi tuntui järkevimmältä. Ei jaksa oikein enää keskittyä.* 'That seemed the most sensible one. It's getting hard to concentrate by now' (Item 27: E7:6)
- *Luulen ymmärtäneeni tekstin, en ihan loppua kyllä* 'I think I understood the text, not the end though' (Item 30: J7:5, B)

There are test-takers who write comments on the procedure of taking the test from their individual point of view, especially in cases where something has turned out to be complicated.

- *För bråttom...* 'Too much in a hurry' (Item 6: H10:8, C)
- *Oikein vaihtoehto on B tai C, laitoin ensiksi B:n joten en viitsi enää vaihtaa*. 'The correct response is B or C, I put B at first, so I don't care to change anymore' (Item 10: S7:6, B)

- *Aluksi en kuunnellut. Kuulin vain jotain esiintyjistä.* 'At first I didn't listen. I just heard something about performers' (Item 25: Z3:3, B)
- *En ehtinyt miettiä A/B mietin liikaa käytetäänkö si vai ei* 'I didn't have the time to think about A/B I thought too much about if si was used or not' (Item 28: J9:6, B)
- *Aioin jo ensimmäisellä laittaa A:n mutta vaihdoin jostain syystä B:hen. B sopisi viimeiseen kommenttiin, muttei sitä edelliseen* 'I was going to put A already at the first time but changed to B for some reason. B would go with the last comment but not with the one before' (Item 28: B9:8, B→)

The test-takers reflect on the part of the text that they have based their selection of options on. It can be noticed that some hesitation has occurred.

- *Ymmärsin vasta kuuntelun loppuvaiheilla, että kyse onkin asunnoista* 'It was only at the end of the text that I understood that it was all about appartments after all' (Item 1: Å6:6, A→)
- *Toisaalta kuulin kyllä mainittavan "he ymmärtävät kyllä jos haluan olla rauhassa"* 'On the other hand I heard him mention "They do understand if I want to be left alone" (Item 3: N12:11, A→)
- *Joihinkin kuultuihin sanoihin perustuva päätelmä* 'An inference based on some heard words' (Item 4: E3:3)
- *Sopi parhaiten…MUTU-perusteella :)* 'The most suitable option...based on a feeling :)'(Item 26: T2:2, B)
- *Tässä oli vähän kahden vaiheilla, mutta kyllä he esittelivät Descartesin ideaa* 'I hesitated between two here, but they did introduce one of Descartes' ideas' (Item 26: K10:6, B→A)
- *Vaihdoin vastausta, koska A sopii sittenkin paremmin* 'I changed my response, since A was better after all' (Item 28: P7:4)

There are test-takers who comment on their feelings towards the test in general or towards separate items, in most cases indicating the complications they have experienced with them.

- *Näin sen ymmärsin. Sanasto on kyllä HUOMATTAVASTI vaikeampaa kuin B2 YO-kokeessa!* 'This is how I understood it. The vocabulary is CONSIDERABLY much harder than at the B2 level matriculation exam!' (Item 4: F4:7, B)
- *Vaikea sanoa, kun monessa kysymyksessä ei meinaa ymmärtää edes vastausvaihtoehtoja. Yritän vastata, mikä tuntuisi ehkä sopivan* 'Hard to say, as in most cases I don't even understand the options. I try to respond with whatever may seem good' (Item 7: F4:7, B) (Item 8: F4:7)
- *Man skulle nästan ha kunnat svara alla alternativ men valde ändå A.* 'Any option could have been selected but I still picked A' (Item 25: D1:2, B→A)
- *En ole varma. Teksti johti harhaan koko ajan* 'I'm not sure. The text kept leading me astray' (Item 26: J1:2 B→)

2LIST: second listening + change of options

An important feature of the listening comprehension items in the test-taking situation is whether the test-takers have only one possibility to listen to the spoken text or whether they are allowed to listen twice. This is very much related to what is included in the construct: what type of information is to be understood by the test-taker, and in what conditions? In some cases it is indeed decisive for the test-takers to get an idea of the general contents of the spoken text during the first listening, in order to be able to focus on detailed or more precise information during the second. With this particular test, therefore, an issue that the test-takers have frequently reflected on has been the possibility to change the response after having listened to the text a second time. In the present research context, the test-takers are asked to leave their original choice of an option, even if they decide to change their choice. This allows an exploration of the different changes of options, as can be seen in Table 32 below:

TABLE 32 Changes of option choices between the distractors and the key after the second listening

ITEM:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Changes of options											
Distr→ Key	1	5	17	14	25	13	21	30	17	15	19
Key → Distr	1	4	14	12	12	8	14	11	19	11	14
Distr → Distr	2	8	17	16	12	10	7	7	13	10	1
Tot	4	17	48	42	49	31	42	48	49	36	34

ITEM:	**25**	**26**	**27**	**28**	**29**	**30**
Changes of options						
Distr→ Key	12	26	11	6	15	13
Key → Distr	13	7	10	8	21	6
Distr→Distr	17	7	16	5	9	3
Tot	42	40	37	29	45	22

It may be especially interesting to consider the influence of the second listening for items 25-30, as these were items at which the text was listened to only once in the original test administration. Would many of the current test-takers have missed on those items if they would not have had the chance to listen to the text a second time? There are, in fact, a total of 93 changes from a distractor to the key for items 25-30 (in approximately 7 % of the total number of responses to these 6 items), showing the necessity in many cases for a second chance to listen. However, in 65 cases (nearly 5 %) the change has been in the opposite direction (see Table 32 above).

According to the information obtained by means of the present test procedure and with the introspection method, in 22 cases the test-takers have been explicit about the advantages of the second listening in items 25-30. In half of these cases the correct response has been found only after the second listening. When adding the examples for items 1-11, we get, for all items, a total of 57 examples where the test-taker has explicitly mentioned the change of the selected option after the second chance to listen (see Table 32). Among these examples there are 31 cases where the correct response has been found after the second listening. Thereby, in nearly half of the cases a second listening has still not helped the test-takers reach a correct response. A Pearson correlation gives at hand that a more difficult item (with a higher item measure) yields significantly more changes from one distractor to another (see Table 33 in Appendix 4).

Sometimes the test-takers reflect explicitly on how the second listening has supported their interpretation or helped them solve the task:

- *Toisella kerralla ehdin kuunnella ja keskittyä paremmin* 'The second time I had the time to listen and concentrate better' (Item 1: E2:2, A→)
- *Tässäkin kohdassa ymmärsin tarinasta vasta II kuuntelun jälkeen…*'Here also I only understood the story after the second listening'(Item 3: P7:4)
- *Ekalla kerralla ei ehtinyt tähän kysymykseen mukaan* 'The first time I didn't catch this first question' (Item 4: B8:8, B→C)
- *Kuulin paremmin toisella kerralla* 'I heard better the second time' (Item 26: K7:5, C→A)
- *Toisen kuuntelukerran aikana selveni enemmän, joten vaihdoin* 'Things became clearer during the second listening, so I changed' (Item 27: R1:2, A→C)
- *Missasin kysymyksen kokonaan ekalla kerralla joten heitin jotain, toisella kerralla selvä juttu* 'I missed the question completely the first time so I just put something, the second time it was a clear case' (Item 29: B9:8, C→A)

There are test-takers who have explained why or how they have changed their option selection after having listened to the text a second time:

- *Ei taidettukaan sanoa, että ne olisi nimenomaan turisteja* 'I suppose after all they didn't say that they were tourists in particular' (Item 3: F4:7, A→B)
- *Puhui laihtumisesta ja vaikeuksista. Aluksi luulin että valot häiritsivät → muutto maalle* 'Talked about losing weight and troubles. First I thought that the lights bothered → a move to the country' (Item 5: U8:7, C→)
- ~~*C oli loogisin, ei ainakaan B*~~., *A kuitenkin, kun naisen olisi pitänyt tietää* ~~'C was the most logic~~, ~~at least not B~~, A after all, as the woman should have known' (Item 28: N10:9, C→)
- *Vastaus onkin B, en I kerralla muistanut, mikä on "ira"* 'The response is B after all, the first time I didn't remember what "ira" was' (Item 29: P7:4, A→)

LOG: Reasoning, based on what is logic

One more or less necessary or inherent strategy of selecting MC options is to reason or think logically about the alternative responses. This inference can be based on for instance world knowledge, experiences or knowledge about the language – referred to as test-wiseness strategies by Cohen 1998a. The most frequent cases of metacognitive responses (with 116 cases, representing one third of the total number) are perhaps therefore the ones where the test-taker gives proof of logic reasoning of some kind. This has been more typical at the pragmatic items where there is some element of reasoning somehow naturally present in the task – the test-takers have to draw on former experiences and knowledge and not simply find the correct option by summarising the text contents.

The essential question is if the strategy of reasoning belongs to the listening comprehension construct. It seems that the top down-processes of drawing on world knowledge or experiences of the language when making decisions on how to interpret messages are justified. In some cases reasoning supports the linguistic interpretation; in others it may compensate for deficient language abilities, which is the case also in real language-use situations. In his study investigating how linguistic and non-linguistic knowledge was activated in a MC test of listening, Yi'an (1998) describes situations where partial success in linguistic processing lead to the activation of general knowledge, either in a compensatory or a dominating manner, which means that a belief was let override even what was correctly abstracted from the text. Yi'an could thus notice the difference in compensating and facilitating functions of the non-linguistic knowledge.

The reasoning evidenced by the present introspective responses is based on world knowledge, classroom experiences, test-wiseness, linguistic features or the text in general. Based on differing degrees of comprehension of the text, the test-taker has sometimes relied on world knowledge to arrive at a conclusion on what the correct response may be. The thirteen cases representing this type of reasoning come from only five items but from 11 different people. This suggests that certain types of items, focusing on a certain type of information, are more likely to call for reliance on world knowledge than others. Among the examples we have the following, six of which are combined with the correct option:

- *Putiikeista puhuttiin…Sitä nyt on tapahtumassa yleensä…*'They talked about boutiques...That is generally happening...' (Item 1: T5:4, A)
- *Puhui, että kiloja putosi, niin on mahd. että se tapahtui sairauksen seurauksena* 'Said that she lost some kilos, so it's possible that this was a consequence of the illness' (Item 5: P8:5, A→)
- *Eihän ne kämpät o koskaan valmiita...* 'The appartments are never quite ready, are they...'(Item 6: L1:5, C)
- *Osti vanhan huoneiston, ei varmasti kallis* 'Bought an old appartment, so it was probably not expensive' (Item 6: K6:5, B→A)
- *Olin kuulevinani jotain kaupungeista, joten ajattelin vastauksen liittyvän matkamuistoihin ☺* 'I thought I heard something about towns, so I thought the response would relate to souvenirs ☺' (Item 6: V7:5)
- *Deras hus var en stor investering så de gör troligen remont, som inte är klar* 'Their house was a big investment, so they're probably doing a renovation that is not finished yet' (Item 6: H8:7, C)
- *Puhuttiin (kait) pöydistä ja ravintoloissa on niitä* '(I guess) they talked about tables and there are tables in restaurants' (Item 10: J2:2, C->B)
- *Hämärä mielikuva, että olisin Pariisissa nähnyt jonkun pienen ruokaputiikin markiisissa épicerie* 'I have an obscure impression of having seen épicerie in Paris on the sunblind of a small foodshop' (Item 10: K20:9)
- *Descartes on kuollut joten on loogista että hänen muistokseen annetaan palkintoja.* 'Descartes is dead so it's logic that they give prices to his remebrance' (Item 26: Z6:4, A→)

Some knowledge-based experiences seem to reflect language learnt in the classroom context:

- *bagarre=tappelu. Jäi mieleen.* 'bagarre=fight. That stayed in my memory' (Item 9: L2:7)
- *L'épicirie on muistaakseni jonkin sortin safkaputiikki* 'L'épicirie was as far as I remember some sort of grocer's' (Item 10: O10:7, A→)
- *"epishri" kuulostaa tutulta. Puhuttu varmaan tunnilla niistä.* '"epishri" sounds familar. We must have talked about them in class' (Item 10: E11:9)
- *När man ber om någonting och måste vänta säger man så* 'When we ask for something and have to wait this is what it is said' (Item 27: H11:10)

Some responses bear on the experiences the test-takers have had of practicing solving MC test items. This practice may have occurred with tests of listening comprehension or of reading comprehension, with French or with other studied languages. The knowledge of what characteristics are typical of key responses or distractors or how the test as a whole should be taken into account when solving individual items - be it called “test-wiseness” or tactics in general - is useful. Rupp (2006) noticed in a study on a MC test of reading comprehension, that the reasoning process through the options is induced by the MC question itself, and thus, unique to the testing context. This relates to the considerations of construct validity and the authenticity of the task – this type of reasoning seems fairly distant from a TLU situation

Interestingly, the 19 cases of this type of reasoning are given in the responses by several different test-takers – it is thus not a case of only a limited few people who make use of this strategy. Another issue concerns the concentration of these responses on specific items. In this test, items 1 and 26 yield the most frequent cases of these responses. The challenge for test constructors is to create items where the key option cannot be found just by being tactical, without understanding the spoken text at all. However, certain test item or option construction methods are typical and perhaps even unavoidable for MC items. Therefore, if a test-taker has had lots of experience with MC items, he or she may recognize these methods and use these clues as a strategy to find the correct option – especially if there is doubt as to the interpretation of the spoken text. The washback effect plays a role here, meaning in practice the relative importance that the individual teacher puts on doing language tasks related to the specific test-format, as opposed to practicing “for life” or for a target language use situation outside the test context.

On the other hand, using all available hints and resources at hand may be considered a useful strategy or tactic for any language learner in any language use situation. The responses show what the test-takers have focused on in these items:

- *Koska tavarataloista ja putiikeista puhutaan liikaa. C tuntui mahdollisimmalta.* ’Because they talked too much about department stores and boutiques. C seemed the most possible one’ (Item 1: S11:6, A→C)
- *Ihmiset asuvat entisissä kaupoissa: seuraava kysymys keskittyy eri asiaan kuin kauppakuolema* ’People live in old shops: the next question focuses on other things than the death of shops’ (Item 1: Å11:7, A→)
- *Valokuvauksesta ja turisteista puhuttiin niin selvästi, että ajattelin, ettei ainakaan ne.* ’They talked so clearly about taking pictures and tourists, that I thought at least not those two’ (Item 4: P9:6)
- *Muut vaihtoehdot varmasti vääriä, koska ne kerran mainittiin niin selvästi nauhalla →hämäys* ’The other options are surely incorrect, since they were mentioned so clearly on the tape → trick’ (Item 9: N8:7)
- *Mainittiin, en tiedä oliko se hämäys* ’It was mentioned, I don’t know if it was a trick’ (Item 10: T9:7, C→B)
- *Eka ajattelin A:ta, mutta se tuntui liian yksinkertaiselta...* ‘First I thought about A, but it felt too simple’ (Item 26: Z14:7)
- *Ei mitään tietoa, ”quitter” yleensä näissä se luottotapaus.* ’I have no idea, ”quitter” is usually the safe case in these’ (Item 27: Å11:7)

Different detailed characteristics of the spoken or the written language itself are considered by some test-takers. They look at the text “from above” as it were - in a top down- manner - not only quoting parts of the text in their responses, but considering it further. Some test-takers rely on the tone of voice or other paralinguistic features to judge the correctness/falseness of the proposed options. Others also focus on syntactic features in the text and in the options, the strategy of which can be paralleled with what Nakatari 2006 calls "scanning strategies". To these types of responses I have also counted the references to the localizations of the necessary information (NI; as judged by the test-takers). There is a proportion of 47 % of correct answers (20 out of 43 cases) indicating that even if this type of processing has sometimes been useful, in some cases it has not been sufficient as a strategy for the test-takers to find the correct option. The test-takers have relied on the tone of voice and paralinguistics mainly for the pragmatic items:

- *Äänensävyn perusteella* 'On the basis of the tone of voice' (Item 28: Q1:4)
- *Se oli pakollista, niin että jos nainen vaikka toistais kun äänensävy oli hiukan sellainen* 'It was compulsory, so maybe the woman would repeat as her tone of voice was a bit like that' (Item 28: Z12:6, B)
- *Jag tycker att både A och C verkar rätta men mannen verkade så säker på sig själv* 'I think both A and C seem correct but the man seemed so certain of himself' (Item 28: H7:7, C)
- *Äänestä/äänensävystä päätellen vastasi näin*. 'Based on the voice/tone of voice she ansered like this' (Item 29: R3:3, A)
- *Hon lät så glad, så det passade bäst med C* 'She sounded so happy, so C was the most suitable one' (Item 30: D2:4)

Syntactic features are focused in the following responses:

- *Jos puhutaan nykyajasta, niin melu ei enää häiritse, ja B-kohta tarkoittaa vissiin tulevaa aikaa, vaivoja sillä oli ennen?* 'If we speak about present time, the noise does not bother anymore, and option B means future time, I suppose, she had troubles before?' (Item 5: B5:6, B>A)
- *Mies kysyy eikö se ole pakollinen, vastaus pakko alkaa si-sanalla…* 'The man asks if it is not compulsory, the answer has to begin with the word si…' (Item 28: K2:3, B)
- *Kielteiseen kysymykseen myönt. vastaus* 'An affirmative response to a negated question' (Item 28: O14:8, B)
- *Oikea aikamuoto* 'Correct tense' (Item 30: U9:8)

Some test-takers have localized the NI:

- *Portti on aina auki ja tauluista ei puhuta,* **alussa**[86] *mainittiin vanhempien pelosta* 'The gate was always open and they do not talk about pictures, **in the beginning** they mentioned the parents' fear' (Item 8: O9:6)
- *Sanottiin* aluksi *inconvenient ja sitten puhuttiin A:sta.* ***Myöhemmin*** *puhuttiin jotain myös B:stä.* **'In the beginning** they said inconvenient and then they talked about A. **Later** they said something about B also' (Item 9: U2:4)
- *Sano* ***lopuks*** *ravintolan joten se ei oo oikein* **'At the end** she said restaurant so it is not correct' (Item 10: T2:2, B→A)
- ***Ensimmäisten lauseiden*** *perusteella* 'On the basis of **the first sentences'** (Item 11: N2:4)

There are several cases of reasoning where test-takers use the clues available from what they have understood of the text to arrive at a choice of an option. Reasoning and inference are considered higher cognitive activities, the question being whether they are compatible with the listening construct. In the cases where the inferring is based on the spoken text, they obviously are. One evident hypothesis would be that the more text has been understood, the more there is

[86] Bold by the researcher

to base the inference on and the more certain the test-taker is of a positive outcome: the selection of a correct option.

- *On kadottanut kiloja. Siitä päättelin, että muutenkin terveydestä valitti* 'Has lost some kilos. From that I concluded that she complained about her health in general' (Item 5: E11:9, C→)
- *Kun jokin kantaa hänen nimeään, niin uskon sen olevan palkinto.*'When something carries his name, I believe it is a prize' (Item 26: J10:7)
- *Ei kai se tyttö sitten tiennyt että se oli pakollista jos ei kerran varannut paikkaa.* 'I guess the girl did not know that it was compulsory if she didn't book a seat' (Item 28: B5:6: B→)
- *Om man nu vill ha filén så svarar man väl C* 'If one wants the filet one would surely answer C' (Item 29: D1:2, A→C)
- *Sopii luontevasti keskustelun jatkoksi.* 'This is a natural way to continue the discussion' (Item 29: Y4:8)

OPT: problems with the question (options or stem)

The cases where test-takers have explicitly mentioned having problems with understanding some of the questions or options are interesting and informative from the point of view of validity as we want to find out if there is a risk of construct-irrelevant variance[87]. If the test-takers have had trouble with an item due to unfamiliar written vocabulary he or she may not be able to pick the correct answer even if the spoken text has been understood (problem mentioned also in the study by Yi'an 1998). Thus a factor (understanding particular written vocabulary or syntax) that is considered irrelevant from the perspective of the listening comprehension construct proper influences the test outcome.

In this test, the most frequent cases of problematic vocabulary are experienced in item 1: as many as 17 test-takers (nearly 8 % of all test-takers) have mentioned that they do not understand a particular option. Even test-takers with high person measures have mentioned problems. The problematic option has been option 1b, and in some cases also 1c has been mentioned:

- *Tuntui parhaalta vaihtoehdolta. En ymmärtänyt muita tarkalleen.* 'Seemed like the best option. I did not understand the others very well' (J3:3, A)
- *En ymmärtänyt muita vaihtoehtoja (B,C) ja mielestäni puhuttiin kauppojen sulkemisesta* 'I didn't understand the other options (B,C) and I think they talked about closing down shops'(E7:6, A)
- *Ymmärsin kyllä mitä nauhalla sanottiin, mutten ymmärrä vastausvaihtoehdoista B & C kohtia, B vahvempi arvaus* 'I did understand what was said on the tape, but I don't understand options B & C, B is a stronger guess' (U7:7)

The other item where difficulties with understanding the question is frequently mentioned (in 14 cases) is item 8. Here it is the key word in the stem that has caused problems:

- *Cambriolages? Nä, ja vet int* 'Cambriolages? No, I don't know' (D1:2)
- *Arvasin, en ymmärrä kysymystä.*'I made a guess, I don't understand the question' (K19:8)

There have been difficulties with vocabulary in the options or the question in the other items as well (see Table 31 in Appendix 4). Of relevance is the way this has influenced the test-takers' test-taking processes and the selection of strategies that have been adopted to solve these problems. Some test-takers have had to ignore the options with un-

[87] This issue is also discussed in earlier chapters where the reasons behind the use of the strategies of guessing and elimination are described.

familiar vocabulary/phrases, and have selected an option that they understand, if only there is some match with the spoken text:

- *Ainut vaihtoehto, jonka ymmärsin* 'The only option that I understood' (Item 4: Z1:2, B)
- *Valitsin B, koska minä ymmärsin sen, muista kohdista en ole ihan varma* 'I picked B, because I understood it, I'm not sure about the other options' (Item 30: B6:7:B)

Sometimes test-takers are left with one option that they have eliminated as incorrect and another that is not understood:

- *Muut vaihtoehdot eivät tuntuneet sopivilta, en kylläkään tunne déranger-verbiä*. 'The other options don't seem good, even if I don't know the verb déranger' (Item 4: Y4:8)
- *En ymmärtänyt A:ta, C kuullosti parhaalta* 'I didn't understand A, C sounded the best' (Item 7: B3:4, C)
- *Se restaurant sana oli hämäyst siin tekstis. En tiedä mitä C meinaa mut ku A ja B ei käy.* 'The word restaurant was a trickery in the text. I don't know what C means but as A and B are not good' (Item 10: L1:5, B→)
- *A se ei voinut olla ja C:tä en kunnolla ymmärtänyt* 'It couldn't be A and I didn't understand C properly' (Item 25: S6:6, B)

Guessing has been the only possible strategy for some test-takers with comprehension difficulties[88]:

- *Arvaus, ei ymmärrä kaikkia vaihtoehtoja* 'A guess, I don't understand all the options' (Item 5: X5:6, C→)
- *Arvasin, koska en tiennyt mitä muut vaihtoehdot tarkoittavat.* 'I made a guess, since I didn't know what the other options mean' (Item 6: J8:6, A)
- *Arvaus (en tiedä sanaa cambriolages)* 'A guess (I don't know the word cambriolages)' (Item 8: U8:7)
- *Ei B, eikä kai C, jota en muuten ymmärtänyt. Veikkasin A:ta*. 'Not B, and probably not C, that I did not understand by the way. I made a guess on A' (Item 9: N10:9)

The test-takers have used all the pieces they have understood in the spoken text – to be compared to what Nakatari 2006 calls "less-active listener"-or "word-oriented" strategies - and have tried to cope with that knowledge, even if the options have caused difficulties:

- *Hon prata någonting om pengar...och så förstod jag inte helt och hållet de andra alternativen* 'She said something about money...and then I didn't understand the other options completely' (1: D2:4, A)
- *Melu häiritsee, en tiedä mikä on santé (totuusko?). No, ainakin puhu jotain provincesta eikä pitäny melusta* 'The noises bother, I don't know what santé is (a truth?). Well, at least she talked about province and didn't like the noise' (5: T2:2, C)
- *Hän pitää pianonsoiton kuuntelemisesta, en ymmärrä kysymystä* 'He likes listening to the piano, I don't understand the question' (9: T5:4)
- *C:stä en saanut selvää, mutta jalkakäytävästä puhuttiin paljon, ja koska ei tuntenut itseään yksinäiseksi, B on mahdoton* 'I couldn't make sense of C, but they did talk a lot about the pavement, and since she did not feel lonely, B is impossible' (11: N8:7)

UNC: uncertainty/problems

The second most common response type classified as metacognitive cover the cases (N: 108, 23 % of the total number) where the test-takers have reflected on

[88] See further ch. 9.3.5 on the strategy of guessing.

the difficulty or uncertainty of understanding or interpreting the text. Of all these cases 40 % is combined with the correct response.

This response type is most frequent for the text passages related to items 3 & 4 and to item 10. There are generally a lot of unclear issues with items 3 and 4, items experienced to be generally difficult. As what comes to item 10, the necessary information lies to a large extent in a key word that has been a source of great confusion. For some items it seems to be the combination of the text and the proposed options that cause difficulties, as in the three cases below:

- *Ainoa kysymys, minkä ymmärsin kokonaan eli ARVAUS* 'Only question that I understood completely that is a GUESS' (Item 1: J2:2, A)
- *Ei mitään hajua mitä se näistä vaihtoehdoista kuvasi*. 'No idea which of these options he described' (Item 3: Z12:6)
- *Svårt, det kan också vara B*. 'Difficult, it can also be B' (Item 6: D1:2, C)

Many responses reflect the fact that the test-takers have been uncertain as to the contents of the text passage; they have understood fragments, but not the necessary information. The strategy then employed varies from guessing in one case to elimination in another.

- *Puhuttiin jotain kaupasta...mutta en ole varma, että mitä tapahtui* 'They talked about a shop... but I'm not certain as to what happened' (Item 1: Z11:5, A)
- *Kadulla tapahtuu jotain, en saanut selvää.* 'There was something happening in the street, I couldn't make it out' (Item 9: Å1:4)
- *Jotain juhlia yöllä, en kuullut muuta* 'Something about parties at night, I didn't hear anything else' (Item 9: F6:11)
- *De talade om ett pris, men jag är inte säker på att det var nytt...(B också möjl.)* 'They talked about a price, but I'm not sure it was a new one... (B also possible)' (Item 26: H12:10)
- *Neiti pyytää häntä pysymään linjalla (en tosin ole ihan varma miten se sanotaan)* 'The lady asks him to hold the line (although I'm not certain how it is said' (Item 27: K10:6)

Some test-takers have simply admitted to not understanding the text. The only available strategy seems to be guessing, even if there may also be cases where the responses give away themselves.

- *Arvasin (en tiennyt yhtään mitä nauhalla puhuttiin)* 'I made a guess (I didn't know at all what they were talking about on the tape' (Item 3: R12:8, C→B)
- *Arvaus, en tajunnut mitään tästä osiosta! sanan sieltä ja täältä.* 'A guess, I didn't understand anything of this item! A word here and there' (Item 4: U8:7)
- *Arvasin, ei mitään käsitystä sanoista.* 'I made a guess, no idea about the words' (Item 7: J8:6)
- *Nauhalla keskeisiä sanoja, joita en ymmärtänyt...* 'There are essential words on the tape that I didn't understand' (Item 10: Z8:5)
- *Päättelin, sillä en ymmärtänyt kaikkea* 'I inferred, since I didn't understand everything' (Item 25: Å15:8, B)
- *En muistanut kaikkien sanojen merkitystä, arvaus* 'I didn't remember the meaning of all the words, I made a guess' (Item 27: B3:4,A)
- *Arvasin (en ymmärtänyt tekstiä/asiaa)* 'I made a guess (I didn't understand the text/subject)' (Item 27: Q1:4, B)
- *En ymmärtänyt kaikkia sanoja* 'I didn't understand all the words' (Item 30: B3:4, A)

There are test-takers who have been able to point to the problem with the spoken text more specifically - the text has not been grasped due to the fact that it has been experienced to be unclear or spoken in a rate that has exceeded the limit for the test-takers. This condition is naturally related to the previous cases: the phonological or segmenting problems lead to the impossible task of even attempting to make an interpretation of the meaning of the text.

- *Arvaus, en kuullut.* 'A guess, I didn't hear' (Item 6: Z8:5, A)
- *? liian epäselvää, arvaus* '? Too unclear, a guess' (Item 9: T9:7, C)
- *Sekava ääni, mutta sain kuitenkin pääteltyä vastauksen* 'A messy voice, but somehow I could infer the response' (Item 10: J1:2, B)
- *En ole varma, sanottiin p:llä alkava sana josta en ole varma onko* 'I'm not sure, they said a word beginning with p of which I'm not sure if it is' (Item 10: J10:7, C→B)
- *Arvasin, en oikein saanut selvää puheesta* 'I made a guess, I couldn't make out what they said' (Item 25: R1:2, B)
- *Puhuivat liian nopeasti, vastaus on puhdas arvaus* 'They talked too fast, the response is a pure guess' (Item 28: J1:2, A->C)

Some test-takers have for some reason missed some parts of the text (the most or all of it) with the obvious consequence that they have to rely on strategies to try to compensate for that circumstance. Depending on the extent of the missing, the strategy is elimination, reasoning or guessing:

- *Ihan silkka arvaus, meni vähän ohi...* 'A complete guess, I missed it a little' (Item 1: V7:5)
- *Gick nog lite förbi...* 'I did miss this a little...' (Item 3: H10:8, B)
- *En osaa sanoa. Kohta meni jotenkin ohi.* 'I can't say. Somehow I missed this passage' (Item 6: S11:6, C)
- *Ohi* ,Missed it' (Item 7: E9:7, B)
- *Mikään ei mielestäni käy. Puhdas arvaus (Meni täysin ohi)* 'None of them is good in my opinion. A pure guess (Missed it completely)' (Item 8: U7:7)
- *Koko kysymys meni totaalisesti ohi!* 'I missed the entire question completely' (Item 27: K2:3, C)
- *Ei hajuakaan* 'No idea whatsoever' (Item 27: K20:9,B→C)

There are test-takers who feel uncertain about their understanding and their interpretation of the spoken text, and/or the combination of this understanding with the proposed options. This is probably related to individual test-takers' characteristics – some test-takers demand a more complete understanding than others in order to feel confident with the task.

- *En ollut yhtään varma, mutta arvasin* 'I wasn't sure at all, but I made a guess' (Item 2: S1:2, B)
- *Tästä en ollut 100% varma, vaihdoin vielä* 'I wasn't 100% sure about this one, so I changed' (Item 4: R1:2, A→C)
- *En tiedä, vastasinko oikein* 'I don't know if I responded correctly' (Item 4: P7:4, C→B)
- *En ole varma, mutta näin luulisin kuulleeni* 'I wasn't sure, but this is what I think I heard' (Item 9: R11:8, C)
- *Nyt ei oo taas varmaa, saatoin vaihtaa väärään…*'Again I'm not sure, I may have changed to the wrong one…' (Item 11: B5:6:B→C)
- *En ollut varma kumpi valituista.* 'I wasn't certain which one of the selected ones' (Item 25: S6:6, B→)
- *Mä luulen et se ois toi, mut en tiedä* 'I think it is that one, but I don't know' (Item 29: U4:5)
- *En ole vieläkään varma, mutta tuntuu siltä, että A* 'I'm still not sure, but I think it's A' (Item 30: B9:8, C→A)

CERT: Certainty

Contrary to the type of responses that reflect uncertainty, there are test-takers who state that they have felt confident when faced with particular items. These responses may reflect the fact that even if the task in general – solving the listening comprehension test items - has been difficult, the test-taker has been positively surprised at individual items that have seemed less complicated. The same test-takers tend to give these kinds of responses; among the test-takers responsible for the 39 responses of this type, test-taker R10:8 (person measure: 59.51) has given six, test-takers E2:2 (person measure: 33.39) and Z8:5 (person measure: 48.73) five, test-takers J1:2 (person measure: 33.39) and Z10:5 (person measure: 48.73) three. This type of reaction seems to be conditioned by the test-taker's personality perhaps more than the actual difficulty of the items or the actual success or skilfulness of the test-takers.

A count gives at hand that 23 cases are correct, implying that for more than one third of the test-takers, it has been a false certainty. This subtype of responses is exemplified below with cases where the test-takers say either that they have understood the text or that they know their choice of option is correct:

- *Ymmärsin* 'I understood' (Item 1: R10:8)
- *Luulen ymmärtäneeni* 'I think I have understood' (Item 11: E6:6)
- *Ymmärsin sentään jotain* 'I did understand at least something' (Item 25: J1:2, B)
- *Luulen että ymmärsin oikein* 'I think I understood correctly' (Item 25: R10:8, B)
- *Ymmärsin ainakin keskustelun pääjuonen* 'At least I understood the main thread of the discussion' (Item 27: Z8:5, C)
- *Muut väärin (nyt tuntui varmalta, ymmärsin asian)* 'The others are wrong (now I felt certain, I understood the issue) (Item 28: S10:6)
- *Helppo saada irti vastaus* 'Easy to get the response' (Item 1: J1:2, A)
- *Tästä olin aika varma.* 'I was pretty certain about this one' (Item 7: E2:2, B)
- *Tää on varmaan ainoo melko saletti tähän asti* 'This must be the only pretty certain one so far' (Item 8: K20:9)
- *Puhdas tieto. Un epicerie on vähän kuin leipomo* 'Pure knowledge. Un epicerie is a bit like a bakery' (Item 10: R8:5, B)
- *Tästäkin mä oon varma (vrt. 8)* 'I'm certain about this one also (cf 8)' (Item 11: K20:9)
- *Tämän osasin varmasti, reagointitehtävät sopivat minulle hyvin. Sain vastauksen I kerralla* 'I know this one for sure, these reaction tasks suit me fine. I got the response at the first listening' (Item 26: P7:4)
- *Varmahko. Tuttu sanonta muualtakin, jos sopi tuohon paikkaan* 'Pretty certain. A familiar expression from elsewhere, that fits in here' (Item 27: J3:3)
- *Tämä on varmasti oikein* 'I'm certain this one is correct' (Item 29: K3:3, A)

?? : Don't know why

In some cases the test-takers find it difficult to point at the reasons for selecting a particular option. They may thus have left the answer box empty or state that they have guessed. However, there are also cases of explicit comments on the inability or difficulty of justifying the selection of a particular option, and these comments are categorised as metacognitive responses. This experienced difficulty is understandable, since the task of stating one's reasons for selecting a particular option during the test event was a completely new way of behaving in a MC test-taking situation for these test-takers. Examples given by different test-takers include:

- *Ei nyt ihan arvaus, mut en osaa perustellakaan…* 'It's not quite a guess, but I can't really justify it either' (Item 1: B5:6)
- *Arvasin, tai en muistanut millä perusteella päädyin vastaukseen* 'I made a guess, or I don't remember on what basis I arrived at this response' (Item 3: P5:3, B), (Item 4: P5:3)
- *En osaa sanoa, arvasin tai päättelin* 'I can't say, I made a guess or inferred' (Item 3: Q2:5, B→)
- *En osaa perustella* 'I don't know how to justify it' (Item 7: Å2:4, C→B) (Item 8: Å2:4)
- *Puoliksi arvaus, ei kunnon perusteluita* 'Half a guess, no proper justifications' (Item 9: B11:8)
- *Mielestäni vaihtoehto B olisi paras, en osaa oikein perustella...* 'I think option B would be the best one, I don't quite know how to justify it…' (Item 25: V7:5, B)
- *Se vain olisi sopiva vastaus. En osaa sitä erityisemmin perustella.* 'It just would be the most suitable response. I don't know how to justify it in any particular way' (Item 30: V2:3)

The metacognitive responses as a whole: quantitative information

In order to obtain a more covering idea of the extent of the metacognitive responses, I will here relate the quantitative representation of the different subtypes to the test-takers' success and the characteristics of the items.

There is no clear relationship between the person measure and the number of instances of the metacognitive introspective responses for an individual test-taker. There is no significant correlation as a result of a Pearson analysis (See Table 25 in Appendix 4). However, when looking at the individual responses, there seems to be some more metacognitive reflection taking place among the stronger test-takers. It can be assumed that these test-takers generally process the text more automatically, and more time and effort can be set aside on the processes of reflecting – self-monitoring and self-evaluation - and naturally also on the added task of writing introspective answers.

Among the subtypes of metacognitive responses there is also proportionally more explicit logic reasoning with test-takers with higher person measures and, on the other hand, there is progressively slightly less uncertainty expressed among these test-takers. For other subcategories, the proportions do not seem significantly different: the mentions of problematic options are practically equally frequent among the responses given by test-takers on different levels. The situation is the same for the question of the helpfulness of the second chance to listen to the text.

The numbers of responses given for the seventeen different items are presented in Table 31. As far as the relationship between quantitative characteristics of the individual items and the metacognitive responses is concerned, there is not a significant correlational relationship between the item measure and the total number of metacognitive responses for an item (See Table 26 in Appendix 4).

However, when considering the relationship between the items and the different subtypes among the metacognitive responses, a Pearson correlation reveals a significant positive correlation between the subtype UNC and the item measure (See Table 34 in Appendix 4). The fact that there is a positive correlation between the number of cases of uncertainty and the item measure seems logic. The higher the value, the more difficult the item and the more probable is the situation where a test-taker feels uncertain about the outcome of the task.

Information on the test-taking process provided by the metacognitive responses

These metacognitive responses reflect the many-faceted and complicated process of solving MC items assessing listening comprehension. Not only is the dimension of grasping, understanding and interpreting the spoken text important, but, added to that, the dimensions of the purpose of the task (defined by the test questions and options) and of the test-takers' various characteristics and their background knowledge also play an important role in the test-taking situation.

If we consider the schematic picture of what the process looks like (see Figure 14 in Appendix 3 on the test-taking processes and strategies), we can notice how these responses reflect different stages or levels in the process. The subtype SIT concerns the entire process, from the onset to the visible result of the process, the response. The subtype OPT comes in at the beginning stage already: the imposed purpose of the task may be blurred, unclear or even misunderstood by the test-taker in case the question and options are not understood. This naturally affects the rest of the stages in the test-taking process in a negative way.

At the stage where the spoken text input is given, the feeling of certainty (CER) or uncertainty (UNC) enters the situation – the test-taker will either feel confident or inconfident about having understood the text, in the light of the task at hand. The second listening (2LIST) have for many test-takers been helpful or necessary in order to be able to grasp the key information in the text.

As for the task of answering the question with one of the options, some logic reasoning (LOG) is often demanded. The combination of bottom-up and top-down processing is the most covering way of solving a task: making use of all the background knowledge useful in the task of interpreting the spoken text in the light of the test question. This interactive process (as described in chapter 1.3) can be evidenced in the present introspective responses.

9.4 How do the test-takers' listening and test-taking processes relate to their success in solving the items?

The introspective responses described and discussed above allow the conclusion that there are differences in the way different test-takers handle the task of selecting an option in different items. The differences are partly determined by the nature of the individual items but are also, besides other individual test-taker characteristics and preferences (see ch. 1.7) related to the test-takers' success level. The assumption is for a test in general that the test-takers' measured ability (as described in the construct) is reflected in the scores on the items. In reality, the situation is more complicated. Depending on the way the text is processed, understood and interpreted, a strategy is selected, and these two stages have an impact on the outcome for the individual test-taker and for the functioning of the item. In fact, all these factors are interrelated. As far as the current test-takers are concerned, their ways of handling a task can be considered and compared as a function of different characteristics. The test-takers' mother tongue (in this case Finnish or Swedish) can be considered, as well as their sex, their school or their home region (urban or rural for instance) for example in determining DIF[89]. Their individual learning or problem-solving styles also have an effect on the processes. However, as this particular test is taken from the school-leaving context, even if for instance the teaching methods between different individual teachers may differ, the test-takers as a group, all upper-secondary students, can be taken to be fairly homogeneous. The majority of the test-takers, as the majority of pupils learning French as a foreign language in Finland on the secondary level, are Finnish-speaking girls living in urban regions. Therefore, within the scope of this study, these potentially influencing factors are not taken into account separately.

The test-takers' ability as defined by the person measure would be a clear quantitative and objective criterion serving as the basis for the comparison in this study. As the person measure is based on a fairly low number of items, however, it cannot be taken as a reliable reflexion of their listening ability, but only as reflecting their success in solving the current seventeen items - a tendency rather than an absolute value or truth. One possible approach to the investigation of the processes and strategies that the test-takers have activated when solving these seventeen items of listening comprehension is the relationship between the person measure of the test-takers and the introspective responses. The correlation of the number of different introspective responses per test-taker and the person measure shows that there are significant relationships between the person measure and certain introspective responses (See Table 25 in Appendix 4).

The most significant correlation is that between the résumé responses and the person measure. The relationship is positive, implying that the higher the person measure, the more résumé responses are observed for the seventeen items. This seems very logic: the more able test-taker understands more of the spoken text and proves this by giving an introspective response containing a summary of the key information in the text – or the other way around: if more of the text is understood, a successful option selection leading to a higher person measure is more likely.

A slightly weaker but a positive correlation is also found with the person measure and the partial comprehension. There are thus more responses partially covering the text contents as a function of a higher person measure.

[89] Differential Item Functioning, defined in the *Multilingual glossary of language testing* terms as "The fact that the relative difficulty of an item is dependent on some characteristic of the group to which it has been administered, such as first language or gender".

For the third response category related to the spoken text, word-bound responses, the relationship is the opposite: there is a negative correlation between the number of word-bound responses and the person measure. The weaker the test-taker, the more frequent is the tendency to rely on single words in responding to the test item. A focus on single words is not a successful strategy.

As far as the strategy of guessing is concerned, a less successful test-taker tends to make more guesses than the more successful test-taker. This is also logic: if a test-taker does not understand anything of the spoken text, the option cannot easily be selected by means of comparing the representation of the spoken text with the options, and there is a greater need to use compensating strategies like guessing.

There are thus differences between the processes and strategies employed by test-takers as a function of the results on these items. In broad lines, while the weaker test-taker struggles with the task of selecting an option on the basis of rather incomplete textual clues, often being inclined to using the strategy of guessing, the stronger test-taker understands the most of the spoken text, and is able to, on the basis of his or her interpretation, eliminate unlikely and impossible options. These tendencies can be compared with the results by Young (1997), Rost (2002) and Vandergrift (2003) on the strategies employed by advanced listeners (See ch. 1.5.3). Skilful inference, more effective elaboration and monitoring in flexible combinations seem to be typical for higher-ability listeners in general.

The differences in success are further due to the different starting points for the test-takers. While the preview of the question and options has been established to be a facilitating factor for advanced test-takers (for example in the study by Yi'an 1998) – providing anticipations and foci for listening - the less able listeners are not able to profit from this circumstance. The MC format seems to favour the advanced listener, adding difficulty to the task for the weaker listener (Yi'an 1998).

It should be pointed out that these main patterns can be distorted when the test-taker is faced with individual items, as a function of some trait – typically a flaw - in that item. In MC tests of listening comprehension, parallel to the situation with reading comprehension, as established by Rupp et al. (2006), it seems obvious, on the basis of these introspective answers that characteristics of the questions and the text interact with characteristics of the test-takers to induce response processes that are mediated by prior experience with such tests.

What the outcome for an individual test-taker is in the end is usually visible for the test administrator or user only in shape of the selection of an option, judged as either correct or incorrect and given out as a final score. The quantitative information is the only trace we have left of the multidimensional process that lies behind this information, if we don't track down the details in the process through some qualitative information. This is what the introspective responses provide. What the results suggest is that there are many elements in the process identified by the introspective responses that cannot be claimed to be found in a TLU – or non-test – situation.

The question is how large a part of the scores for a test-taker is affected by construct-irrelevant factors, what the consequences for the reliability of the test-scores and the validity of their use are, and whether and how these factors could be eliminated in a test-situation.

In the following chapter, I will focus on separate items in the light of what the introspective responses reveal of their functioning. Judged by the quantitative analysis, the items worked as expected by the Rasch model, matching the level of the test-takers. However, the numbers reveal only a part of the truth. The test-takers raise issues in their responses that indicate some problems in the items.

9.5 Some problematic features of the individual items

Contrary to what is suggested by the experienced constructors of the present test items (see chapter 4.2), it has not proved easy to predict exactly how the MC items functions. What is maintained by for example Alderson (2004) or Buck (2001) seems to hold true: it is difficult to write MC items and be able to foresee how they work in an assessment setting. The introspective responses to the seventeen MC items indicate the presence of some problems in the items. In some cases the reasons behind these problems seem more easily detectable than in others. If the reason behind the complexity or the obscurity of an item can be relatively easily detected, it is probably possible to revise the item slightly so that the test-taker is more likely to be faced with a test item where he or she can show his or her listening comprehension ability. In that case, the scores obtained are likely to become valid and reliable. An overview of the items where some problems seem to occur, at least for this group of test-takers that are taken to be representative of the target group for the current test items, are given in Table 35 below. In the following, I will describe and discuss the problems in the items in the light of the introspective responses as well as the content analysis and the details in the statistical information. It is important to consider validity issues on the item level, as pointed out by Haladyna (2004: 262).

TABLE 35 Problems in different items reflected in the introspective responses

Item	Type of problem	
1	Difficult vocabulary in options	
3	Random guessing	
6	Random guessing Order of the items not following the order of information in the text	
7	Random guessing and nonsense responses combined with the key; Difficult vocabulary in options	Heavy processing load due to many options (9) to be treated at the same time
8	Difficult vocabulary in question	Order of the items not following the order of information in the text
9	Difficult vocabulary in question	
11	(All processes and strategies lead to the selection of the key)– Relatively implaus-	

	ible distractors
25	Difficult vocabulary in options
27	New task type; Difficult vocabulary in options: need to recognize particular phraseology
29	Implausible option Two correct options?
30	Implausible option

Item 1

The quantitative summary of the types of introspective responses that item 1 yields (see Figure 25) gives at hand that metacognitive comments and guesses are the two most typical response types. There are very many changes from one option to another between the two chances to listen (see Table 32 in chapter 9.3.7). This is probably partly due to the fact that this is the first item for the test-takers – both the test format and the theme of the text are foreign for them.

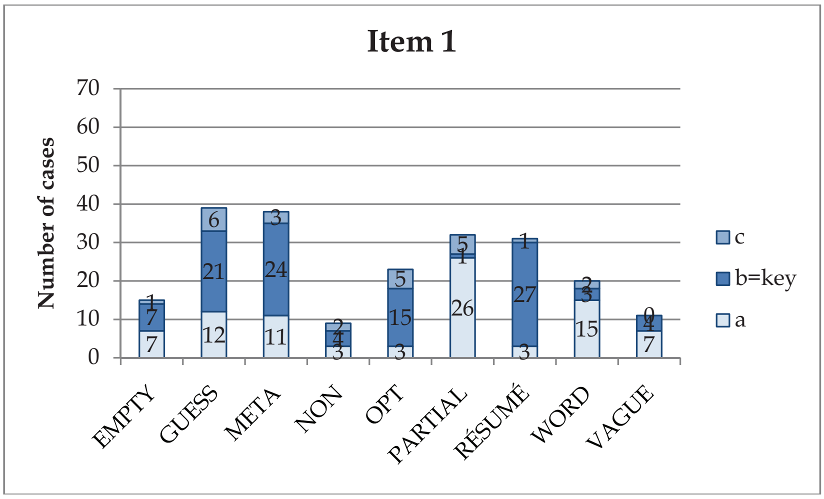


FIGURE 24 Item 1: introspective responses combined with the three options

Nearly two thirds (63 %) of the metacognitive responses are given with the selection of the key option. Slightly more than the half of the guesses are made on the correct option. Interestingly, partial comprehension is evidenced very clearly in combination with distractor 1a (see further chapter 9.3.2). The problem with this item is related to the situation where the options have not been understood: as large a proportion as 8 % of the test-takers indicate that they have experienced problems with options 1b and 1c. This would serve as a counterargument against the use of this item in its current shape. Among the comments that indicate a non-understanding of the options there are the following[90]

- *No en oikeastaan ollu varma mitä noi vaihtoehdot on suomeksi, mutta mun mielestä siinä puhuttiin hintatason noususta*. 'Well I actually was not sure about what those options are in Finnish, but I thought they talked about the rise in the price level'(Z7:5)
- *En ymmärrä paljoakaan C:stä mutta ei A tai B oikeen kuulosta oikeilta* 'I don't understand much about C but neither A nor B sound really correct' (U4:5)
- *Ei ole ainakaan a eikä c (en ymmärrä b:tä)* 'At least it is not a nor c (I don't understand b)'*(N12:11)*
- *En tiedä mitä rez-de-chaussée tarkoittaa* 'I don't know what rez-de-chaussée means (T1:1),
- *En tiedä kaikkia vastausvaihtoehtojen sanoja, A liittyy ainakin jotenkin asiaan* 'I don't know all the words in the options, A has at least something to do with the issue' (J7:5),
- *En ymmärtänyt muita vaihtoehtoja (B,C) ja mielestäni puhuttiin kauppojen sulkemisesta !* 'I didn't understand the other options (B, C) and I thought they talked about closing down shops!' (E7:6),
- *Kuulosti parhaimmalta vaihtoehdolta. En ymmärtänyt B ja C vaihtoehtoja* 'Sounded like the best option. I didn't understand the options B and C' (S10:6).
- *En kyllä oikeen ymmärtänyt → 1B ja C eli arvasin* ' I didn't quite understand → 1b and c so I made a guess' (N9:8)

Based on these responses, the options should be rewritten to make them more transparent and fair for the test-takers.

Item 3

Quantitatively, item 3 is the least conform to the Rasch model. This can to a large extent probably be due to the strong tendency to guess at the item, as evidenced by the introspective responses (see Figure 26). All in all, item 3 seems to have been a difficult and problematic item because of the processing load of the compact text, and the complicated task of matching the inferred message of the text contents with the key option. The question is whether the item can be judged to be too demanding for test-takers at the target level for the test. In that case, there is construct-irrelevant variance due to the test contents not matching the contents of the construct. This may potentially affect the test-takers: they may be confused, with negative consequences for the test-taking processes as well as for the outcome. Some indeces of unfamiliar expressions in the options are also given and these add to the complications.

Here are examples of responses to item 3, describing the test-taking procedures including indications of problems with understanding either the text or the options:

- *Pelkkä arvaus, kuuntelu meni ohi* 'Just a guess, I missed the spoken text' (E2:2, B)
- *En ymmärtänyt yhtään…* 'I didn't understand this at all…' (V7:5, B)
- *Gick nog lite förbi…* 'I missed this a little bit…' (H10:8, B)
- *Tämä meni molemmilla kerroilla vähän ohi. Piti ainakin ulkona olemisesta* 'I missed this a bit on both listenings. At least he liked to be outdoors' (K10:6)
- *En oikein ymmärtänyt → arvasin* 'I didn't quite understand → I made a guess' (O11:7)

[90] See also related examples on introspective responses with indicated comprehension problems in chapters 9.3.5, 9.3.6 and 9.3.7.

- *En ymmärtänyt B:tä (vaihtoehtoa) mutta A ja C eivät sopineet* 'I didn't understand B (the option) but A and C weren't good' (Q4:8, B)
- *En ihan ymmärtänyt vaihtoehtoja, mutta B vaikutti loogisimmalta.* 'I didn't quite understand the options, but B seemed the most logic one' (N10:9, B)
- *Arvaus. En edes ymmärtänyt mitä C tarkoittaa* 'A guess. I didn't even understand what C means' (B7:5)
- *Ei mitään hajua mitä se näistä vaihtoehdoista kuvasi* 'No idea whatsoever which of these options he described' (Z12:6)
- *Hän kai puhui työstään, joten vaihdoin vastauksen* 'I guess he talked about his work, so I changed the answer' (S1:2, B)
- *Sopi toisen kerran jälkeen vaan paremmin, ei varmaa tietoa.* 'Was simply more suitable after the second listening, no certain knowledge' (S6:6, B)
- *Uppfattade inte "arbetslösa" på första gången, därför B på andra gången* 'Didn't catch "unemployed" at the first listening, therefore B at the second' (H8:7, B)
- *En lukenut kysymyksiä aluksi, koska luulin, että 1 & 2 tulevat heti uudestaan ja toisella kerralla kirjoitin edellisen perusteluja. Eli ohi meni* 'First I didn't read the questions because I thought 1& 2 would come directly a second time and on the second time I wrote justifications for the last one. So I missed this.' (Z3:3, B)

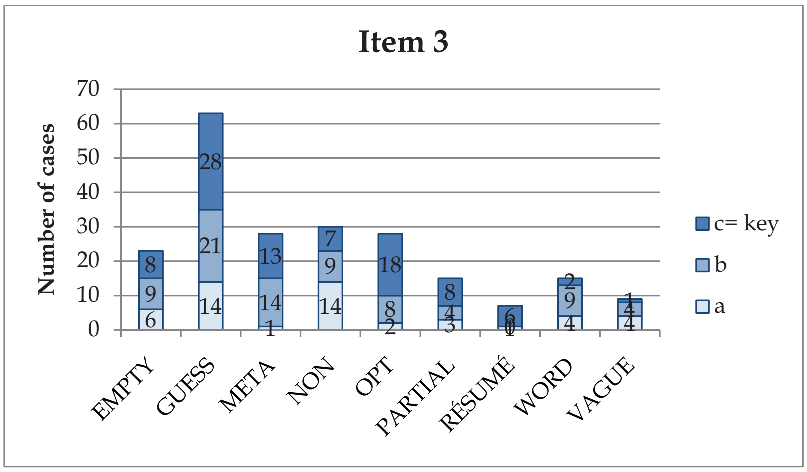


FIGURE 25 Item 3: introspective responses combined with the three options

An attempt should be made to simplify the text and perhaps make the options more transparent. However, the functioning of the revised item remains unclear until it has been tested on another group of test-takers.

Item 6

For item 6 likewise, the strategy of guessing has been reported to have been frequently used (see Figure 26). As contrasted with item 3, however, the discrimination has been stronger. The tendency is more clearly that stronger test-takers – even if they state that they have made a guess - have selected the key and weaker test-takers have ended on a distractor. Similarly to item 3, however, the text has been experienced to be demanding, and one of the options has not always been understood.

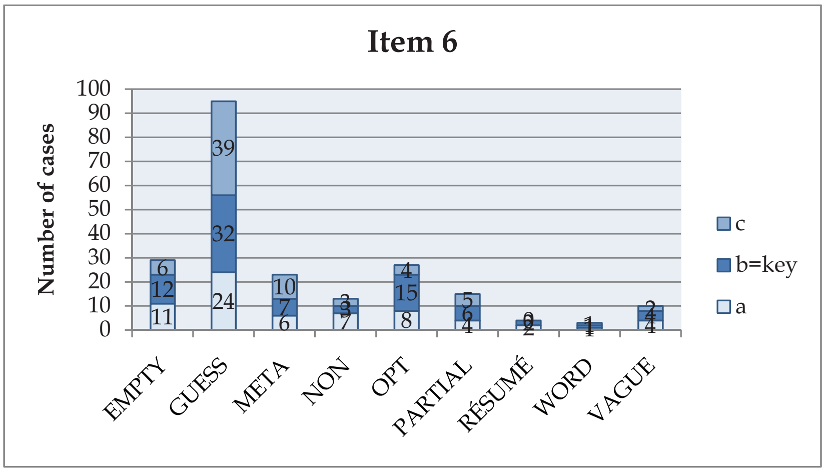


FIGURE 26 Item 6: Introspective responses combined with the three options

- *En saanut kovin hyvin selvää, enkä ihan ymmärtänyt täysin varsinkaan kohtaa kuusi* 'I didn't grasp this very well, and didn't understand completely especially item 6.'(P7:4, A)
- *En saanut selvää paljoakaan*. 'I didn't catch very much.'(Å9:7, A)
- *Taloudellisia ongelmia, hämärää…* 'Economic problems, obscure…' (E8:7)
- *En kuullut mainittavan mitään a tai c:hen liittyvää, joten otin b:n, jota en ymmärrä* 'I didn't hear mentioned anything related to a or to c, so I took b, that I don't understand (N12:11)
- *Svårt, det kan också vara b* 'Difficult, it can also be B' (D1:2, C)
- *Asunto oli kallis, enkä ymmärtänyt b-kohtaa* 'The apartment was expensive, and I didn't understand option B' (V5:4, C)
- *En osaa sanoa, kohta meni jotenkin ohi* 'I can't say. I missed this item somehow' (S11:6, C)
- *prêt= ?* (E11:9, C)

Item 6 is testing, among other things, the ability to compare negated – i.e. relatively more demanding - written statements with the text content, and the attitudes of the speaker towards a situation, expressed mainly verbally, but also by the tone of voice. To this adds the circumstance of the necessary information for item 6 comes already in the beginning of the passage, before the necessary information for item 5. This is obviously experienced to be complicated.

Guessing is generally used more often as a response strategy when the options are opaque, that is unclear, not unambiguously wrong or correct, or if they contain unfamiliar expressions. There is obviously something unclear with the options for item 6, since there are very few text-related responses and a vast majority of instances of the use of the strategy of guessing. This tendency may be reduced by changing the text, the order of the text (or the questions) and the options slightly.

Items 7, 8 and 9

The combination of three items 7, 8 and 9 to be treated at the same time, with the questions not following the order of the dense text contents has caused problems for the test-takers, as evidenced by different responses and response patterns in the introspection. Added to that, one of the options in item 7 has been considered difficult. For both items 8 and 9, the stems include difficult vocabulary.

For item 7, the most frequent response is guessing (see Figure 27). There are very few responses related to the spoken text, that is, of the type résumé or partial comprehension.

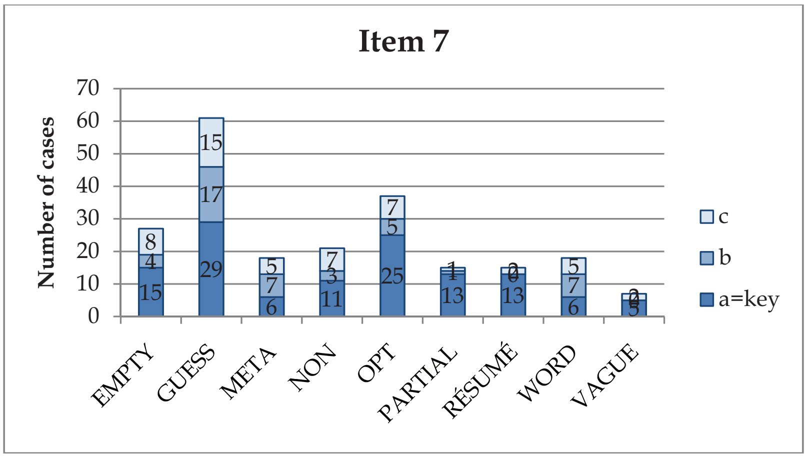


FIGURE 27 Item 7: Introspective responses combined with the three options

The difficulties in this item may be due to unfamiliar words or expressions in the options:

- *Vaikea sanoa, kun monessa kysymyksessä ei ymmärrä edes vastausvaihtoehtoja. Yritän vastata, mikä tuntuisi ehkä sopivan* 'Difficult to say, as at many questions I don't even understand the options. I try to answer what might suit' *(F4:7)*
- *Ymmärsin vain A kohdan* 'I only understood option A' (B6:7)
- *En ymmärtänyt kysymyksiä, paitsi C:n ja päädyin A:han* 'I didn't understand the questions, except C, and I ended up with A' (N10:9)

Some test-takers indicate that they have missed the text or the task:

- *Meni hieman ohi, arvaus* 'I missed that, a guess' (E6:6)
- *En oikeen kuullut...joten arvasin* 'I didn't really catch that…so I made a guess' (F5:9)
- *Arvasin, ei mitään käsitystä sanoista* 'I made a guess, I didn't have a clue about the words' (J8:6)
- *Piti arvata A:n ja C:n välillä. C mielestäni todennäköisemmin oikein. (Meni täysin ohi*) ' I had to guess between A and C. I think C is more probably correct. (I missed it completely). (U7:7, A→C)

The help or necessity of the second listening is reflected in many cases:

- *Ensimmäisellä kerralla en ymmärtänyt mitään, toisella sain kiinni ideasta. Vastaukset sen sijaan ovat tod.näk. päin mäntyä. Arvasin ne kuulemani perusteella* 'At the first listening I didn't understand anything, at the second I caught the idea. The responses are probably all wrong. I guessed on the basis of what I heard' (Items 7, 8 and 9: P7:4)
- *Heillä on ovi usein auki (?), "un jardin" – käsitin merkityksen vasta 2.kerralla* 'They often have the door open (?), "un jardin" – I only understood the significance at the 2. time' (K16:7, B→)
- *Kuulin paremmin* 'I heard better' (R10:8, B→C)

At item 8, as a contrast to items 3, 6, and 7 for example, there is a larger proportion of text-based responses and fewer guesses (see Figure 28).

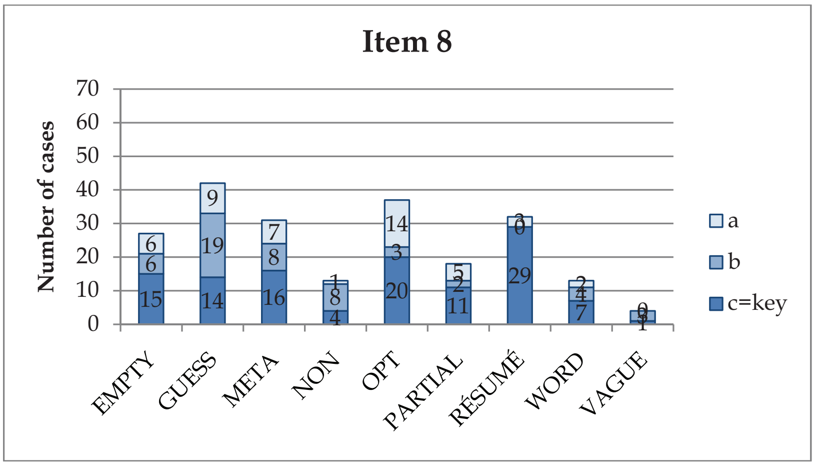


FIGURE 28 Item 8: Introspective responses combined with the three options

However, there are examples of responses to item 8 indicating problems with understanding the question or the task:

- *Mitähän toi kysymys tarkoittaa*? ' I wonder what the question means?' (B2:2, B)
- *Cambriolages? Nä, ja vet int* ' Cambriolages? No, I don't know' (D1:2)
- *Arvasin, en ymmärrä kysymystä.* ' I made a guess, I don't understand the question' (K19:8)

- *Ainoa vaihtoehto jota en ymmärtänyt, ja muita ei mainittu tuolla tavoin* ' This was the only option that I didn't understand, and the others weren't mentioned like that' (S9:6, A)
- *En ymmärrä kysymyksen sanaa, joten arvasin ☺* I don't understand the word in the question, so I made a guess ☺' (Å8:7)
- *(hämärä)* '- (obscure)' (Å11:7, B→)

Therefore, even if item 8 has functioned generally well, the piece of vocabulary that seems to be unfamiliar to many test-takers may have caused invalid processing for some. If reused, the item should be altered with respect to that detail – an alteration that seems easily made.

There is some unclearness as to item 9. It seems to have been difficult to select the key based on a simple correct interpretation of the text. Much guessing is called for (see Figure 29), not least due to a misunderstanding of or a lack of focus on the stem: many test-takers have not observed that a negative characteristic is asked for. Consequently, it has been difficult to be able to select the pieces of information in the text that should or should not be included in the comparison with the options.

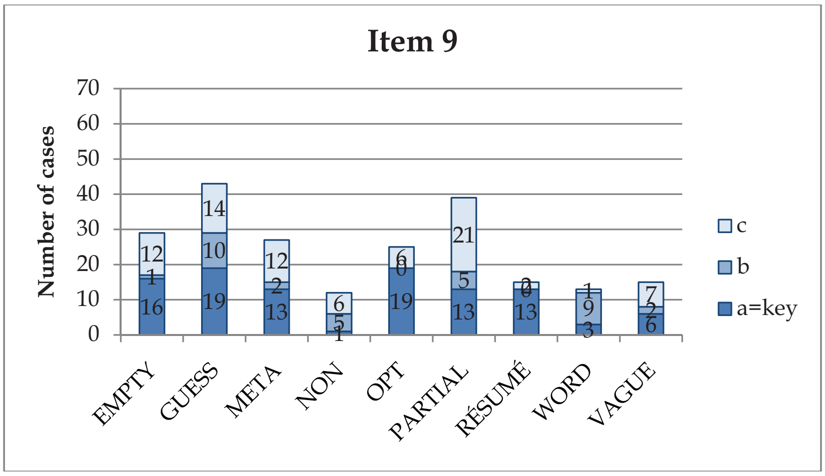


FIGURE 29 Item 9: introspective responses combined with the three options

The comments given by the test-takers on this item concern the test-taking procedure, as

- the lack of time:
    - *Ööh, en tiedä, vähän meni arvauksen puolelle, kun ei kerkinyt lukea vastauksia kunnolla* 'Uh, I don't know, I had to guess a bit, since I didn't have the time to read trough the options properly' (T1:1, C)
- the lack of attention or understanding
    - *Arvaus. En kuunnellut* 'A guess. I didn't listen'(Z3:3, C)
    - *Arvasin. (Tämä osio meni ohi)* 'I made a guess (I missed this item)'(Z7:4, C)
    - *Meni ohi.... Tuntui parhaimmalta* 'I missed it... Seems to be the best option' (E9:7,)
    - *? liian epäselvää, arvaus* '? too unclear, a guess'(T9:7, C)
- problems with understanding the question
    - *Hän pitää pianonsoiton kuuntelemisesta, en ymmärrä kysymystä* 'He likes listening to the playing of the piano, I don't understand the question' (T5:4)
    - *Arvaus (Inconvenient?)* 'A guess (Inconvenient?)(B6:7, B)
    - *Joskus kadulla on rauhatonta. Frapper=taistella/uhata?!?* 'Sometimes there are disturbances in the street. Frapper = fight/threaten?!?' (Y3:8, C→)
    - *Arvasin...sans frapper...?Ainoa joka tuntu oikealta* ' I made a guess...sans frapper...? The only one that seemed correct' (N9:8, A→C)
- the second listening
    - *Ekalla kerralla kohta meni ohi. Toisella kerralla ymmärsin.* 'At the first listening I missed the passage. At the second I understood' (E1:2, A→C)
- test-taking tactics
    - *Muut vaihtoehdot varmasti vääriä, koska ne kerran mainittiin niin selvästi nauhalla →hämäys* 'The other options are surely wrong, as they were mentioned so clearly on the tape → a bluff' (N8:7)

The task of managing three items combined with one text passage during one pause is difficult. This can be evidenced for example by the the fact that justifications for item 7 contain text contents that have nothing to do with that item but with items 8 and 9 (see chapter 9.3.2 on partial comprehension). Added to that, item 8 is the only item where there are two cases of abandon of the task of selecting an option. It seems that the format of having three items to handle at a time puts undue stress on the test-taker, and should be changed. Moreover, it may be profitable and just for the test-takers to change the order of the items, so that they follow the order of the contents in the spoken text. As the spoken input has to, by nature, be linear, without the chance for the test-takers of moving back and forth in the text (contrary to what is the case with reading) it seems fair that the questions should not deviate from this linearity on the target level of the current test.

Item 11

There should be a small amount of easy items within a pool of items used for a specific test, both for the psychological reason of not putting down weaker test-takers, but also in order to discriminate in the weaker end of the scale of test-taker performance. However, the fact that an item is very easy for a specific group of test-takers may not be an intended characteristic of an item. The item may be testing something that is below the targeted level of the test-takers, so that even the weaker test-takers are expected to manage the task, or then the distractors are so implausible that not even the weaker test-takers select them even in case of an incomplete understanding of the spoken text. The introspective responses reveal some of the underlying reasons for the easiness of item 11.

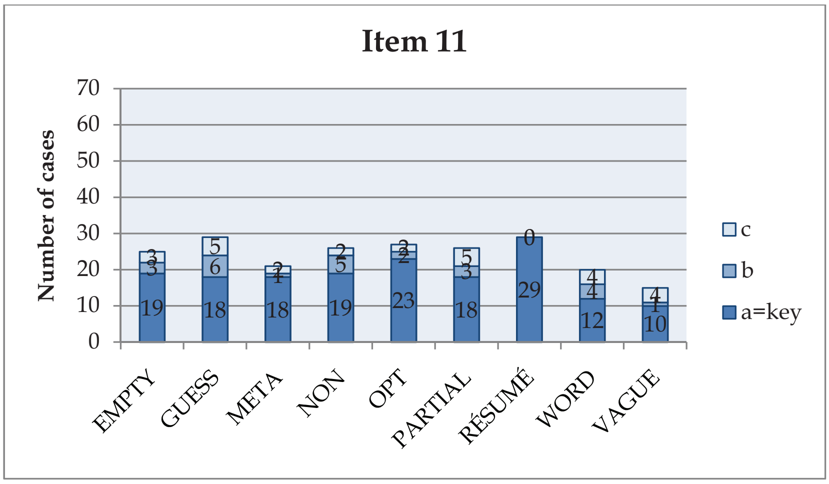


FIGURE 30 Item 11: introspective responses combined with the three options

The pattern in Figure 30 showing the relationship between the introspective response type and the selection of the key option or one of the distractors in item 11 clearly visualises the fact that the use of any response strategy has lead to a majority of correct responses. Interestingly and alarmingly, even the majority (73 %) of the misinterpretations as evidenced by the nonsense responses have ended on the correct option, with examples like:

- *Äiti pyysi häntä usein tekemään kaikenlaista, mutta leikki mieluiten kadulla koiranpentunsa kanssa* 'Mother often asked her to do something, but she rather played in the street with her puppy' (K10:6)
- *Arvasin. Sitä paitsi sanottiin vain ettei se asu yhdessä äitinsä kanssa. Ei mitään siitä näkeekö hän äitiään koskaan* 'I made a guess. What is more, they only said that she doesn't live with her mother, nothing about if she ever sees her mother' (U6:7)
- *Hän oli usein ulkona ystäviensä kanssa, joita hänellä oli paljon* 'She was often out with her friends, which she had a lot of' (Å9:7)
- *Hon hade inget annat problem än att hon ibland var arg på sin mamma för att hon var enda barnet. Annars lycklig barndom.* 'She didn't have any other problems than the fact that she was sometimes mad at her mother for being the only child. Otherwise she had a happy childhood' (H12:11)

One source of misinterpretation is found in the polysemous *"jouer"* 'to play' that some test-takers have interpreted as being connected to playing music:

- *Hänen äiti vaati aina häneltä jotain, kadulla soittaessaan tyttö löytää rauhan* 'Her mother always demanded something of her, playing in the street the girl finds peace' (X1:3)
- *Hän vietti lapsuutensa soittamalla välillä kadulla äitinsä kanssa, ja piti sitä hauskana* 'She spent her childhood sometimes playing music in the street with her mother, and found it to be fun' (Å6:6)
- *Hän oli muiden ihmisten keskipisteenä soittaessaan* 'She was at the center of attention of other people when she was playing music' (Å12:7)

Some test-takers' responses show that even if they have had trouble understanding the text, they have succeeded in reaching the correct option:

- *Nyt ei oo taas varmaa, saatoin vaihtaa väärään...* 'Again I'm not sure, I might have changed to a wrong one…' (B5:6)
- *Meni ohi eli arvaus* 'I missed this so it was a guess' (X5:6)
- *Arvaus, ei ymmärtänyt mitään* 'A guess, I didn't understand anything' (B12:8)

The main problem with this item lies in the reasons for its facility: the distractors do not seem to have been sufficiently efficient. The fact that a weak adverbial expression *souvent* 'often' is used in the key, compared with the two stronger expressions *pas de* 'no' and *jamais* 'never' in the distractors may have given undue help of interpreting the text in light of the distractors, with a favouring of the weaker adverbial. Specific determiners are something that should be avoided, and the importance of the parallellism between the key and the distractors is emphasized by Haladyna (2004) and Mendelsohn & Rubin (1995). Added to this, the negative implications of the two distractors are clearly conflicting if the expression *enfance superbe* 'wonderful childhood' in the text is grasped. It is not clear from the introspective responses that this is how the test-takers have reasoned, but test-wiseness of this kind does help the test-taker in arriving at the correct option in a MC test. In brief, almost any process and strategy has lead to the test-takers choosing a correct option for item 11, which serves as a counterargument against its use in its current form. The question is if a better discrimination would be possible to obtain by a reformulation of the options in order to make them more plausible.

Item 25

The metacognitive comments given in combination with the selection of the three options for item 25 reveal the difficulties faced by the test-takers and the possible problems related to the validity of this particular item. Several themes seem to surface. The first is the helpfulness of the second listening. This is particularly understandable for an item that is of a new and different type both for the test-takers in general, and within this particular pool of items. It is interesting from the point of view that in the original test administration, the text passage is listened to only once (see further Table 32 and discussion in chapter 9.3.7). The test-takers responsible for the responses below have, even if they have had a second chance to listen, still not arrived at the correct option:

- *Tuntui järkevimmältä, ekalla kerralla meni ohi…*'Seemed to be the smartest one, I missed the first time…'(E6:6, A)
- *Ekalla kerralla vain arvasin. Toisella kerralla muutin vastausta, koska se sopi paremmin.* 'At the first listening I just guessed. At the second I changed the answer, since it fitted in better' (E1:2, C→B)
- *Tuntui paremmalta vaihtoehdolta toisella kuuntelukerralla* 'Felt like a better option at the second listening' (Z2:3, A→B)

The second theme is simply the difficulty that the test-takers have experienced in grasping or understanding the text:

- *En ymmärtänyt oikein, mistä kysymys, mutta taidemuseosta* ' I didn't quite understand what it was all about, but it was about an art museum' (S1:2, B)
- *Arvasin, en oikein saanut selvää puheesta* 'I made a guess, I couldn't make out the speech' (R1:2, B)
- *En kuullut aivan tarkasti, arvasin osittain* 'I didn't hear properly, I partly made a guess' (F5:9, B)
- *Ei varma vastaus* 'Not a certain response' (B11:8)

There are also rather confident test-takers who have, nevertheless, all but one, picked distractor 25b:

- *Ymmärsin sentään jotain* 'At least I understood something' (J1:2, B)
- *Olen asiasta MELKEIN varma* 'I'm ALMOST certain about this' (Z10:5, B)
- *Selkeästi vastaukseen viittaava lause* 'There was clearly a sentence that related to the answer' (Å4:5, B)
- *Luulen että ymmärsin oikein* 'I think I understood properly' (R10:8, B)
- *Ymmärsin C:n oikeaksi...ilman arvailuja! ...väärässäkin voin toki olla.* 'I understood that C is correct… without guessing!…Of course I might be wrong.'(N2:4)

As a third theme, the use of test-taking tactics based on the analysis of the way the options are constructed is evidenced in the following two responses. This illustrates the demands of the test-taking process and the interplay between the text and the item with the stem and options:

- *Epäilen, koska nauhalla mainittiin sama verbi kuin B-vaihtoehdossa, mutta ehkä silti paras* 'I doubt it, since the same verb was mentioned as in option B, but maybe still the best one' (N8:7, C→B)
- *Tämä sopi mielestäni parhaiten. A ja B kohtien väittämistä ei puhuttu niin kuin ne väittämissä ovat vaan eri asioita mainittiin. Toiv. ymmärsit* 'I think this was the most suitable. They didn't talk about the statements A and B as they are in the statement but different things were mentioned. Hope you understood' (U7:7)

A fourth theme, and the most important from the point of view of the validity of this item, is in the alarming responses by those who describe problems in understanding the options. In this case the key option 25c has been difficult:

- *Ainoa vaihtoehto, joka tuntui käyvän. (Vikaa en tajunnut)* 'The only option that seemed to suit. (I didn't understand the last one)' (P4:4, B)
- *Ensimmäinen ei sovi, en tiedä mitä C tarkoittaa* 'The first one is unsuitable, I don't know what C means' (K7:5, B)
- *A se ei voinut olla ja C:tä en kunnolla ymmärtänyt* 'It couldn't be A and I didn't understand C properly' (S6:6, B)

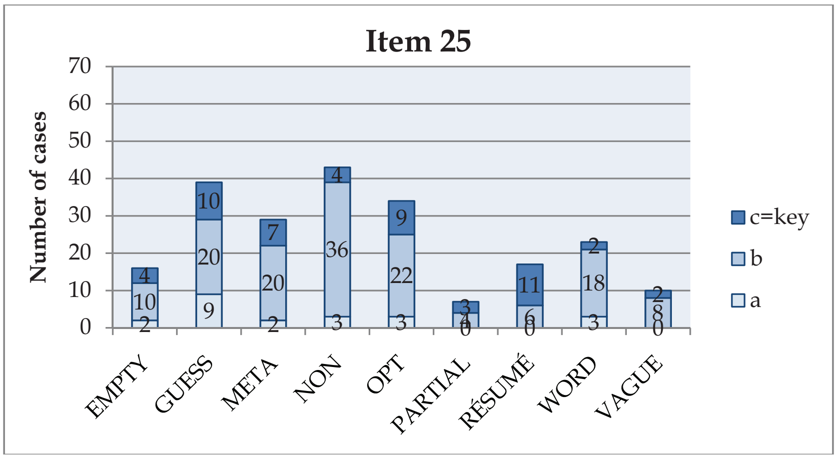


FIGURE 31 Item 25: introspective responses combined with the three options

Item 25 is the item with the largest difference between the small proportion of test-takers having selected the key and the large group of test-takers having selected the most attractive distractor. The question is then how come this item has been so very difficult. One explanation may be the fact that a different type of item (compared with the first part of the test where the text was longer and divided into shorter passages) is introduced. Item 25 constitutes an independent whole. For this research context, item 25 comes after a completely different part of the test with open-ended questions (described in chapter 5.2.1). This may add to the difficulty of the item. However, in the original administration of the test this also turned out to be the most difficult item.

The question is naturally if the attraction of distractor 25b can be a threat to the quality or the validity of the item? This mainly depends on two characteristics: the degree of falseness of the attractive distractor and the clarity of the options in general. The choices of 25a have mainly come about by means of a guess. Many test-takers have thus known to rule out 25a, but have hesitated between the two remaining options:

- *Ei tietoa, veikkasin C:tä. Ei ainakaan A. ajattelin B:tä jo ennen kuuntelua, mutta ehkä kuitenkin C.* 'No idea, I made a guess on C. At least not A. I thougt about B already before the listening, but maybe still B' (N10:9)
- *Inte A i alla fall..eventuellt B men...* 'At least not A...possibly B, but...'(H10:8)

As for the falseness of option 25b, at the very limit, *"des artistes peu connus"* 'unknown artists' in the option could be interpreted as being true according to the text. It demands some inference and integrating of "*art contemporain*" 'contemporary art' "*non-initiés*" 'outsiders', "*rencontres avec les artistes*" 'meetings with the artists'. There has been a source of confusion concerning the target group of the gallery and the focus of the text: the artists or the audience. There is probably an influence from distractor 25b, selected in the following cases:

- *Galleria, jossa näytteillä tuntemattomampia töitä* 'A gallery where more unknown pieces of art are exhibited' (O4:3, B)
- *Taidegalleriasta puhuttiin sekä taiteilijoista jotka ovat katsojille uusia* 'They talked about an art gallery and artists that are new for the audience' (E4:4, B)
- *Paikalla on esillä vähemmän tunettujen taiteilijoiden töitä.* 'Less famous artists' pieces of work are exhibited there' (Å5:5, B)
- *Se yrittää herättää ihmisten mielenkiinnon uusiin artisteihin.* 'It tries to raise people's interest for new artists' (Å2:4, B)
- *Antaa uusia taidevinkkejä yleisölle.* 'Gives new tips of art for the audience' (Y3:8, B)

However, the last noun phrase in the spoken text *"un nouveau type de publique*" 'a new kind of audience' should serve as evidence for the fact that what is focused in the text is the audience, not the artists. Examples of correct interpretations of the text are given by test-takers with relatively high person measures:

- *"on onnistunut hankkimaan uudentyyppisen yleisön"* → *tuntui sopivalta* '"Has succeeded in getting a new kind of audience" → seemed suitable' (Å11:7) (Person measure: 62.76)
- *He ovat avanneet uuden, ohikulkijoille suunnatun näyttelyn* 'They have opened a new exhibition intended for passers-by' (Q5:8) (Person measure: 62.76)
- *Galleria on saanut tavoitettua uuden yleisötyypin* 'The gallery has succeeded in reaching a new kind of audience' (F6:11, B→) (Person measure: 79.41)
- *Taide saa uuden yleisön* 'Art gets a new audience' (O14:8, A→) (Person measure: 59.51)
- *Se houkuttelee kaikkia ihmisiä/ohikulkevia* 'It attracts everybody / all passers-by' (N5:6) (Person measure: 56.60)

There does not seem to be a clear case of two options being true, which would be an unacceptable characteristics of the item. As is pointed out by Alderson et al. (1995), the most important characteristics of a key is that it is unambiguously correct, while the distractors need to be unambiguosly wrong.

The other important and related trait is the clarity and transparency of the options. There are some problems with the comprehension of the key option 25c, which may imply that test-takers who have understood the spoken text are still not able to prove this comprehension by selecting the correct option, if they do not understand its meaning. Ruling-out may in that case be their only alternative. However, it seems, judging by comments given to several items, that the strategy of selecting an option that is not understood is not particularly popular. Here the test constructor or the user of the results of the listening comprehension items cannot be certain whether a key option or a distractor is selected based on a situation where a) the spoken text is understood but the options are not, b) neither is understood, c) the text is not understood, but the options are. The result obtained for a test based on such items does thus not give reliable information on a test-taker's listening comprehension abilities. A revision procedure for this item would be to rewrite the options for more clarity.

Item 27

Item 27 represents a case similar to item 25: a distractor is more attractive than the key. Item 27 is difficult and the estimated discrimination is relatively low.

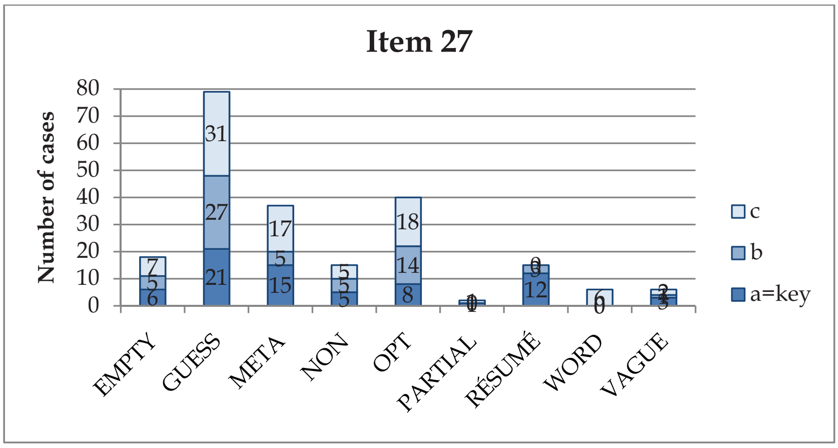


FIGURE 32 Item 27 introspective responses combined with the three options

In item 27, there is very little variation among the types of given introspective responses (see Figure 32). Apart from a few cases of résumés among the correct responses, the guesses largely dominate the image, both for the selections of the key and the two distractors. The key option for item 27 is selected by means of guessing or a comprehension of the necessary information, the latter evidenced by the résumés given as reasons for the choice of the key option. Some test-takers have understood the important context of the phone conversation, but there is still a noticeable uncertainty reflected in many of the responses[91]:

- *Puhelimessa puhuttiin,* ***oletettavasti*** *pyytää olemaan sulkematta puhelinta* 'They're on the phone, she **presumably** asks him not to hang up' (E4:4)
- *Soittaja haluaa jotakin, vastaaja* ***varmaan*** *tarkoittaa, että odota hetki, älä lähde* 'The caller wants something, the answerer **probably** means that wait a minute, don't leave' (T5:4, C→)
- ***(jos)*** *puhu puhelimessa A sopivin* **(if)** he was on the phone A is the best one' (U1:3)

[91] The elements of uncertainty in the responses are made bold by the researcher:

- *Kuulostaa puhelinkeskustelulta, ja A **kai** tarkoittaa, ettei toinen saa laskea luuria* 'Sounds like a phone conversation, and A **probably** means that the other shouldn't hang up' (N8:7)
- *Puhelu, quitter **voisi** olla sopiva verbi* 'A phone call, quitter **might** be a suitable verb' (J12:10)
- *Vastaaja pyytää olemaan linjalla* 'The answerer asks to hold the line' (Å12:7)

The most covering summary of the text contents (and the only response mentioning the concept of 'private accounts') is found in the response by the most successful of all 218 test-takers, J13:11 (person measure: 92.02):

- *Pyytää päästä puhumaan henkilön kanssa joka hoitaa yksityistilejä → joten "pysykää linjoilla" ei muut* 'Asks to talk with the person in charge of private accounts → so "hold on", not the others' (J13:11)

There are also cases of misunderstanding combined with the key option. There are those who have determined that the speakers are both physically present:

- *Arvaus, nainen kai lähti ja pyytää miestä odottamaan kunnes palaa* 'A guess, the woman probably left and asks the man to wait until she comes back' (T4:4, B→)
- *Ajattelin, että haetaan "älä lähde", koska hän pyytää ko. hlön paikalle* 'I thought that "don't leave" is looked for, since she asks the person in question to come there' (E6:6, C→)

Letters, messages and persons are mentioned in the other responses:

- *Kirjeentuojan tulisi odottaa allekirjoitusta* 'The person who brought the letter should wait for the signature' (K12:6)
- *Menee etsimään viestiä (?!). Arvaus…*'Goes to look for the message (?!). A guess…'(Y3:8, C→)
- *Han frågade om hon kunde ta reda på vilka personer saken gällde* 'He asked if she could find out what persons the matter concerned' (H12:10)

There are test-takers who say explicitly that they know this is a phone conversation but that they don't know how "Hold on" is expressed in French:

- *Neiti pyytää häntä pysymään linjalla (en tosin ole ihan varma miten se sanotaan)* 'The lady asks him to hold the line (even if I'm not quite sure how it is said)' (K10:6)
- *Todennäköisempi puhelinkeskustelussa* 'More probable in a phone conversation' (E3:3, C→B)

There are various kinds of problems experienced either in the context of the test-taking situation, or with the text or the task:

The text:

- *Arvasin, koska en ymmärtänyt.* 'I made a guess, because I didn't understand' (E1:2)
- *Arvasin…meni täysin ohi molemmilla kerroilla* 'I made aguess...I missed it completely both times' (S3:4, B)
- *Aika arvauspohjalta. En kuullut hyvin.* 'Basically a guess. I didn't hear properly' (J10:7, C)

The test-taking situation:

- *Arvaus aluksi. En kuunnellut taaskaan* 'A guess at first. Once again I didn't listen' (Z3:3, C)
- *Eka kerta meni ihan ohi, se ukko pyysi jotain, tuntuu sopivalta vastaukselta* 'The first time I missed it all, the guy asked for something, seems like a good response' (T1:1, A→B)

The task:

- *En osaa perustella* 'I can't explain' (P5:3, B)

- *Kaikki tuntuivat oudoilta, joten valitsin sellaisen jonka joku saattaisi sanoa* 'They all seemed weird, so I picked one that somebody might say' (S9:6, B→)
- *Voiko tällaiseen muutakin kuin arvata? ☺* 'Is it possible to do anything else than guess to one of these? ☺' (Z12:6)

Some vague explanations of why the key option has been selected are also given:

- *Sen mitä ymmärsin, niin että virkailija voisi sanoa. Tai siis...äh.* 'What I understood the employee could say. Or I mean…oh, bah.' (V9:6)
- *Ei mitään tietoa, "quitter" yleensä näissä se luottotapaus.* 'No idea, usually "quitter" is the safe solution in these' (Å11:7)
- *Tuntui lähinnä loogiselta vaikken itse olisi vastannut samalla tavalla* 'It seemed the most logic even if I hadn't answered like that myself' (Å14:8)

Item 27 represents a problematic case in that it is testing other things besides pure listening comprehension: "telephone conversation phraseology". The fundamental problem is that in most cases the students have hardly practiced this kind of pragmatic or conversational language use – at least within the genre of "service encounters" - in class at all at the stage of preparing for the test in early 2002.[92] It is obviously a useful function for TLU situations, for any phone conversation task. The question is if this is a content that is included in the construct in 2002? Are the test-takers expected to be able to realize that the two distractors are clearly wrong, 27b syntactically and 27c semantically?

There are consequently at least two types of hypothetical situations that may distort the reliability of the scores on this item:

1) A situation where the test-takers who understand the text but not the pragmatic or idiomatic value of (some of) the options have to make a guess that may result in the selection of a distractor.
2) A situation where the test-takers who do not understand the contents of the text but make a correct guess - possibly based on knowledge about common phone conversation expressions.

The question remains if there is something that should or could be changed with this item.

Item 29

The problems associated with item 29 relate to the circumstance of one distractor that is not clearly and unambiguously false, while one distractor is implausible. There are only some marginal cases of selections of distractor 29c, arrived at mainly by means of guessing mainly by test-takers with lower person measures.

[92] Today, with the inclusion of the type of more pragmatic or conversational items in the test during the past years, the researcher assumes that this kind of language is practiced more commonly in class.

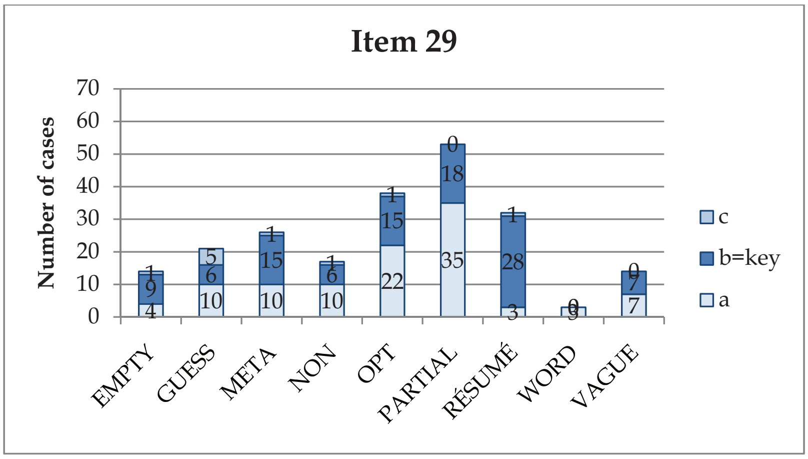


FIGURE 33 Item 29 introspective responses combined with the three options

The decisive single piece of vocabulary in the spoken text that gives a clear reason to select the key comes at the end of the first speech line. Only a few test-takers, mainly, but not exclusively, from the top end of the scale of person measure have grasped the fact that the meat to be bought is not intended for the woman but for the cat, and these test-takers have selected the key:

- *Det handlade om vad katten tyckte så därför B* 'It was about what the cat liked, therefore B' (H2:4) (Person measure: 37.24)
- *Se on kissalle* → *A ei käy. Hän ottaisi filettä* → *C ei käy* 'It is for the cat → A is not good. She would take filet → C is not good' (Å2:4) (Person measure: 48.73)
- *Maten var till katten, det fanns ingen filé* 'The food was for the cat, there wasn't any filet' (H7:7) (Person measure: 51.29)
- *Nainen ajattelee, että entrecôte kelpaa kissalle fileen sijaan* 'the woman thinks that the entrecôte would do for the cat instead of the filet' (O9:6) (Person measure: 53.89)
- *Vastaa tähän: Otatteko entrecôteeta? Liha tulee kissalle.* 'Answers this: Would you have some entrecôte? the meat is for the cat' (S12:7) (Person measure: 59.51)
- *Looginen jatko, mikäli filettä ei olekaan enää kissalle.* 'A logic follow-up, if there isn't any filet for the cat after all' (Å14:8) (Person measure: 66.60)
- *Halusi kissalleen filettä, mutta se oli loppunut, joten A ei sovi, B ei sovi.* 'She wanted filet for her cat, but it was finished, so A isn't good, B isn't good' (N12:11) (Person measure: 79.41)

The category of partial comprehension is large (see Figure 33) . The four pieces of information in the text that have been understood are: 1) the woman wants some filet, 2) there is no filet, 3) something else is offered instead and 4) the woman does not want it – which is in most cases inferred from the woman's speech line saying she "naturally" wants filet. Many test-takers have understood these pieces of information and have selected distractor 29 a.

The ruling-out strategy is frequently made use of, reflecting the situation where the test-takers know that 29c can be ruled out, the final selection thus being between 29a and 29b. Many changes of options have occurred from the key 29b to the distractor 29a:

- *C ei ollut oikein, arvaus An ja Bn välillä* ’C was not correct, a guess between A and B’(B3:4, A)
- *Mielestäni A on sittenkin ainut mikä kävisi miehen kommenttiin. C se ei ollut ainakaan...* ’I think A would after all be the only possible option to the man’s commentary. At least it wasn’t C’ (V7:5, A)
- *C ei käynyt ja B ei sovi keskustelun aiheeseen yhtään*. ’C wasn’t good and B doesn’t go with the subject of the conversation at all’ (S11:6, A)
- *Joo, arvaus A:n ja B.n välillä* ’ Yeah, a guess between A and B’ (K20:9, A)
- *B ja C eivät sovi tarjoilijan kysymykseen.* ’ B and C don’t go with the waiter’s question’ (Z6:4, A)
- *Kysymys → kielteinen, C ei sovi ja B ei sovi* tilanteeseen ’The question → negated, C isn’t good and B does not go with the situation’ (S8:6, A)
- ~~*B, ei C ainaskaan, A oli toinen mahdollinen vaihtoehto, mutta päädyin B:hen*~~ *A kuitenkin, kun naiselle ehdotetaan jotain muuta fileen tilalle...* ’~~B, at least not C, A was the other possible option, but I ended on B~~ A after all, as they suggest something else instead of the filet…’(N10:9, A)

Several test-takers point out the fact that both 29a and 29b would do:

- *En nyt tiedä, tässähän käy A ja B, jos nyt oikein vaihtoehdot ymmärrän suomeksi* ‚Well I dont know, I think both A and B would do, if I understand the options in Finnish‘ (V6:5)
- *A vaihtoehto olisi myös käynyt, mutta koska mies ehdotti, niin siihen voisi vastata d’accord* , Option A would have been ok as well, but since the man suggested, the answer could be d’accord‘ (J5:4)
- *Ajattelin, että nainen vastaisi ehdotukseen, että kai se sopii. A:kin olisi ehkä käynyt, ei C.* ‚I thought the woman would answer the suggestion with I guess it’s ok. A might have been ok too, not C‘ (Z7:4)
- *Sopii parhaiten (en kyllä tiedä onko se liian suorasti sanottu) B:kin se saattaa olla* ’This was the best option (even if I don’t know if it’s too directly put) it might also be B‘ (U7:7)

Some test-takers reflect on the conversational or socio-linguistic aspects of the situation:

- *B, kohteliain vastaus* ’ B, the most polite answer’ (Q3:7)
- *De hade inte "filet" och han frågade om hon kan ta entrecôte i stället. Artigare att svara med B än C ☺’* They didn’t have "filet" and he asked if she would take entrecote instead. It is more polite to answer B rather than C ☺’ (H12:10)

There are inferences based on different textual or non-textual information. Some test-takers consider the facts in the text and the natural speech turns in a conversation:

- *Ku ei o filettä ni se kai vois kelail mitä se sit ottaa* ’As ther’s no filet so I guess she could think about what to take instead’ (L1:5)
- *Näin täytyy reagoida* ’This is the way to react’ (R11:8)
- *Filettä ei ole enää, vastaus A olisi epäkohtelias* ’There’s no more filet, the answer A would be impolite’ (T7:4)

Other test-takers reflect on the tone of voice used by the speakers:

- *Jaa, ei se henkilö ollut ainakaan* tyytyväinen ’Well at least the person wasn’t happy’ (T1:1)
- *Äänestä/äänensävystä päätellen vastasi näin.* ’ Answered this on the basis of his (tone of) voice’ (R3:3)

The validity of item 29 would be improved by changing two options: creating two clearly false but plausible distractors. This would avoid the causing of confusion and would make the task fairer for the test-takers. The challenge for item writers is obviously to find or create oral texts to which it is possible to write good questions and options.

Item 30

The introspective responses reveal that the seemingly easy item 30 owes part of its facility to the fact that the key 30c has been selected because distractor 30b has seemed implausible and because the vocabulary in distractor 30a has been considered difficult.

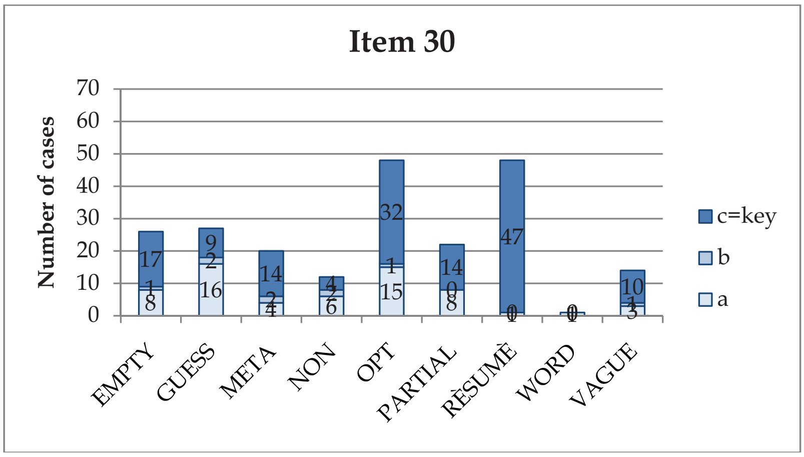


FIGURE 34 Item 30 introspective responses combined with the three options

The option-focused responses (see Figure 34) reveal how the test-takers have felt certain about the incorrectness of option 30b:

- *A:sta ja C:stä tuntui paremmalta* 'Of options A and C this seemed better'(Z6:4)
- *Parempi?* 'Better?' (U3:5, A→)
- *Ei ainakaan B joten jäi 2 valittavaa. Päädyin sittenkin C:hen.* 'At least not B so I was left with 2 to choose from' (V7:5, A→)
- *C oli paras, ei B, enkä usko että A:kaan. C on aika varma veikkaus, toivottavasti oli oikein... ☺* 'C is the best one, not B, and I don't believe in A either. C is quite a certain guess, hopefully it's correct ☺' (N10:9)
- *A ja B eivät ole loogisia, C kai menettelee.* 'A and B are illogical, C is ok I guess' (Å14:8)
- *A mennyt muoto, C voisi sopia..* 'A past tense, C could be correct…'(J12:10)

The responses reflecting difficulties in understanding one of the distractors are:

- *En ymmärtänyt kaikkia sanoja* 'I didn't understand all the words' (B3:4 :A)
- *Valitsin B, koska minä ymmärsin sen, muista kohdista en ole ihan varma'*I chose B because I understood it, I'm not sure about the other options' (B6:7, B)
- *Suljin taas pois vaihtoehtoja, yhtä en ymmärtänyt edes.* 'Again I ruled out options, one of them I didn't even understand' (Å8:7)

- *En ymmärtänyt a:ta, ja arvoin sitten b:n ja c:n väliltä* 'I didn't understand a, so I made a guess between b and c'(V5:4)
- *A tai C, en kyl tiedä mitä chouette tarkoittaa, eli ehkä sittenkin C…*'A or C, even if I don't know what chouette means, so maybe C after all…' (B5:6, A→)
- *B ei käy. A:ta en ymmärrä joten C'* B is not good. I don't understand A, so C' (K19:8)

In order to avoid construct-irrelevant variance, the distractors for item 30 should be made more plausible without potentially difficult vocabulary.

Conclusions on the problematic item features

As a summary, there are some issues emerging from the analysis of the introspective responses with a focus on problems encountered at individual items. First, there is the situation where the processing of the text combined with the task (stem and options) seems to cause excessive demands on the test-takers. Examples of demanding items are item 3, item 6 and item 25, plus the combination of the three items 7, 8 and 9. The nature of the text with a high information load thus seems to be an important factor (See the information processing model by Jamieson et al. 2000 and the concept of cognitive load by Brown 1995). The question is if some of the items are too complicated for the average test-taker at the target level for the test, and even so for the strongest ones among the test-takers, who should be successful even when faced with a more demanding item. An unduly difficult item incites random guessing, which may distort the test results. The results obtained in a study by Rupp et al. (2006), having compared the reading behaviour in test and non-test situations indicate that the response processes to individual items is a complex process with certain abstractable common features with the main factor being the perceived difficulty of the text or the questions. He found that if an item is perceived as more difficult, there is a continual back and forth between the question and relevant text sections in order to logically eliminate potentially incorrect choices. This process continues until a potentially correct option is confirmed or until fewer options remain and the final choice is done by guessing. The similar processes are likely to operate with listening comprehension – even if the interaction between the question, options and the text has to be altered to function between the options and the mental representation of the text, which makes the process even more complicated and demanding for the test-takers.

The implausible distractor is a problem in cases where it has been ruled out also by test-takers who have not understood very much of the text (a possible situation with items 11, 29 and 30). For these test-takers a random guess between the two other options is more likely to be a lucky correct guess than in the case where the guess has to be done among three options. The idea with the distractors is that they represent possible and likely misinterpretations of the spoken text, so that test-takers who do not understand the key information in the text select a distractor (cf. the recommendations on the nature of the distractors given by Ebel & Frisbie 1991; Haladyna 2004 and Linn & Miller 2005).

Another problem is represented by the case where an option intended as a distractor is not sufficiently clearly wrong according to the spoken text (exemplified in item 29). In such a case, the distractor may be selected by test-takers who have understood the spoken text, which obviously leads to construct-irrelevant variance in test scores.

There is a serious problem linked to the situation where the opacity of the options have restrained the test-takers who have understood the spoken text from being able to prove their understanding by means of the selection of the key option. This is a danger Ebel & Frisbie (1991) warn against and seems to be a serious issue related to the typical flaws in the nature of a MC item: vocabulary or expressions in the question or in the response options that are not understood by the targeted test-takers. The results by Yi'an (1998) in his study on the effect of the MC format on the test-takers' performance showed that misinterpretations of options lead to incorrect answers whereas test-takers' uninformed random guessing risked leading to correct answers for wrong reasons.

In some items in the test under scrutiny (items 1, 7, 8, 9, 25 and 27), test-takers have explicitly mentioned having had problems with understanding the options or the question. This may lead to different types of behaviour. Many test-takers seem reluctant to select an option that they do not understand. Others may use their test-wiseness and take the use of infrequent vocabulary in an option as a characteristic related to the key option. There are probably test-takers who give up an item in which they find unclear vocabulary. From the point of view of the quality of an item, this implies possible construct-irrelevant variance, where a score may be missed not as a consequence of incomprehension of the spoken text, but as a consequence of incomprehension of a written sentence or a word.

In many of these items containing a problematic feature the processes and strategies activated by the test-takers are affected by this feature. There is, as judged by the introspective responses, a shift from the focus on the spoken text – which is expected to be the essential focus in a test of listening comprehension – to strategic processing. A focus on the options (elimination) and guesses are frequent. This can be illustrated as in Figure 35, where a focus in the centre of the construct – the spoken text – is what is intended. With different features and problems in the test situation, the focus is moved outwards and away from the construct.

A feature related to the research context that has to be kept in mind here is that, as a contrary to the original test situation, the questions (the stem and options) are read through and can be reflected on twice. In the original test situation, the test-takers first listen to the text in its entirety before reading through and answering the questions while listening a second time (items 1-11) or read through the questions and answer them after having listened only once (items 25-30) (see chapters 5.2.1 and 5.2.2). This may, as a consequence, add to the focus on the questions in the current research context.

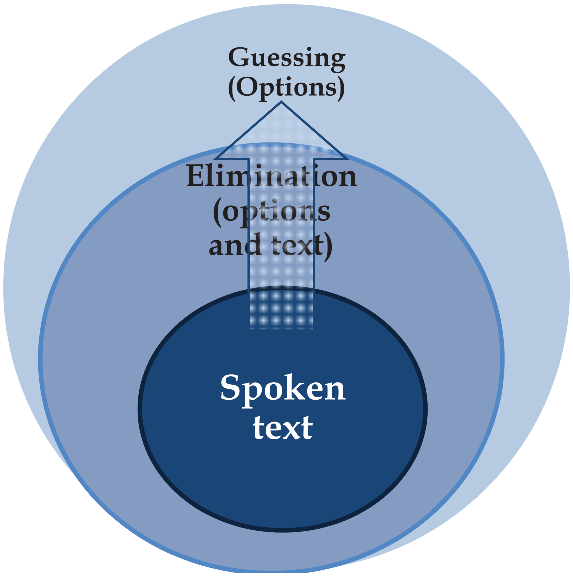


FIGURE 35 The focus in the test-taking process from the text to the test-taking strategies

In terms of the possible revision of the present items some changes to the items seem more easily feasible than others. A difficult detail in an option or a question or the order of the questions seems to be easily changeable. On the other hand, it may be difficult to pinpoint the complications in the spoken text, as they may be a result of the interaction of several features, with or without an influence from the test task. A more thorough think-aloud procedure would be needed to get deeper into the exact reasons behind the difficulties in the text experienced by the test-takers, or a trying out of changes of different potential sources of difficulty at the pre-testing stage.

In the last part of the thesis, some conclusions related to the research questions are drawn and discussed further (chapters 10.1-10.3). The limitations of the research paradigm and its practical implementation are treated in chapter 11. In chapter 12, potential implications and applications of the current research study for the listening comprehension process, for the item format, for the validation procedure, as well as the assessment and L2 learning context are discussed. Some propositions for further research studies are given in the final chapter 13.

10 CONCLUSIONS AND DISCUSSION

Based on the puzzle work of the analysis of the listening comprehension items from the point of view of validity and reliability, I will draw some summarizing conclusions in the following chapters. The focus is on the text and the items, on the results on the items, on the test-takers and on their ways of responding to the task from all possible angles.

10.1 What processes are activated and what strategies are employed by the test-takers on seventeen multiple-choice items assessing listening comprehension of French as a foreign language?

I hypothesized that there would be clear influences from the test format on both the automatic processes and the deliberately chosen strategies when the test-takers are facing a test of listening comprehension in the multiple-choice format. The method effect on the processing and on the test outcome is evidenced also for example in the studies on a MC tests of listening comprehension by Yi'an (1998) as well as in studies on MC test of reading comprehension by Rupp et al. (2006). Based on the introspective responses, taken as a whole, a wide spectrum of processes and strategies seem to be used. Some of the responses are related to the listening process as such, visible through the focus on the spoken text, either on elements in the text, ranging from separate words to larger chunks or summaries of the necessary information, or on experienced difficulties of grasping or understanding the text or of interpreting the main message in the context. The listening process, taken as a whole and judging by the introspective responses, seems to follow a heterarchical, cooperative (cf. Greene 1986) or a parallel or interactive model (cf. Flowerdew & Miller 2005).

As far as the word-bound responses are concerned, both the proportion of these responses and the subtype of response are related to the results of the test-taker on the current seventeen items: the higher the person measure, the fewer word-bound responses are given altogether. The reliance on single words as a strategy is also mentioned in the Oral Communication Strategy Inventory (Nakatari 2006). In the present study, the weaker test-takers differ from the stronger test-takers in their tendency to focus primarily on words present in the options that they try to match with the spoken text. On the other hand, the stronger test-takers who do give responses of this type tend to give responses reflecting words found in the text and not in the options.

For the responses evidencing partial comprehension of the text the average success rate of 47.5 % suggests that skilful inferencing and elaboration have sometimes made success possible despite of incomplete comprehension – supposedly combined with a successful employment of different strategies. For some items, the combination of the test-taker's partial comprehension with the focus on some traits of the item has thus been sufficient to arrive at a correct option selection. Rost (2002; see also Haastrup 1987) mentions the strategy associated with successful listening where incomplete information is combined with inferencing skills to handle a demanding situation.

The fact that 94 % of the résumé-responses (for all 17 items) are combined with a selection of a correct option is expected. In an optimal case, for a valid and reliable test, the proportion should reach 100 % as a sign that test-takers who have understood the essential message of a spoken text have found the key. This is probably seldom possible in reality due to for example different disturbing factors in the test situation, representing random sources of variance to the test scores.

Test-takers who have misinterpreted the text (as proved by the nonsense responses) have often combined some piece of information in the text with the contents of one of the options. Sometimes the piece of text fits the key option, which is thus selected by means of an "invalid" strategy. Contrary to the situation where the test-takers have understood the main message of the text, in cases where they have not, in a "valid" test situation we would expect that they would give an erroneous response to the task. This is, however, not always happening when the MC format is used.

The test-takers react in different ways to the task. Some show affective reactions, whereas others give proof of a general capacity of metacognitive processing, where the test-taking situation is monitored. A general tendency is that more metacognitive reflection takes place the higher the scores of the test-taker (see Young 1997 and Vandergrift 2003). This may also be taken as a sign that as less effort is put into the text processing, there is simply more capacity left for reflections - as well as for the task of writing these reflections down.

Both test-takers' feeling of desperation of their own deficiencies in the situation and their capacity of analysing the basis of their decision-making are visible through their introspective responses. Uncertainty is clearly reflected relatively more frequently for items that are either more difficult (according to the item measure) or for some reason unclear.

Judging by frequent introspective responses, we can infer that the test-takers' focus of attention is to a large extent on other parts of the task than on the understanding of the spoken text. The question - stem and options - impose a purpose for listening, and a lot of effort is put on decrypting the meaning of these four lines of written text. This parallels what is pointed out also by Buck (2001), Brindley (1998) and Alderson et al. (1995). The underlying fundamental task for the test-taker is to form a mental representation of the spoken text, on the basis of whatever is grasped and understood of the text, and compare this representation with each of the three options, in order to determine their truthfulness in light of the spoken text (see the description of the test-taking process by Jamieson et al. in Table 7, chapter 2.4.2). Due to the item format, different strategies can be, and are indeed frequently, employed to cover for, for instance, insufficient comprehension or as a backup because of a feeling of uncertainty. The characteristics of the test-takers represent an important factor, i.e. their level of confidence and certainty when faced with a particular task, most likely influenced by earlier experiences of similar tests (cf. Rupp et al. 2006). According to studies described by Rantanen (2003: 187) for example, the feeling of uncertainty has an effect on the tendency to guess on a MC test task. Many learners as well as their teachers engaged in the present research study mention the anxiety experienced especially with the listening comprehension test component of the assessment of foreign languages in the Matriculation Examination. This is partly due to the transient nature of the language source: there is no possibility of moving back and forth between the spoken text and the question. A test-taker should be confident enough to rely on his or her ability and the individual mental representation of the text, even if neither may be experienced as being covering and complete.

The most typical strategies, found in the context of the selection task format, are guessing and elimination (see Haladyna 2004). The presence of these strategies generally causes harsh criticism against the MC format. The fact that guessing is used does not only reveal something about the nature of the MC test format in general, but about the quality of the particular items that a particular test consists of. Faced with a difficult text, a test-taker is more likely to have to rely on guessing if he or she has not grasped enough of the text to base a confident answer on. The questions and the options have a strong influence on the strategy: if the question is difficult to understand, if the task demanded is too difficult for the targeted level or if the written language is too complicated, thus causing the task to become opaque, the test-taker’s only solution may be to rely on guessing (see Yi’an 1998).

There are clear problems associated with the strategy of guessing, whereby both the validity of the item use and the reliability of the test scores can be questioned – as a result of the fact that random guessing hardly is a construct-relevant ability and may lead to a correct choice and unreliable test scores. However, guessing can also be anything but random (see Linn & Miller 2005; Rupp et al. 2006; Bachman & Palmer 1996). Informed guessing, whereby a test-taker relies on all the clues he or she has based on the spoken text, perhaps combined with reasoning, seems to be related with what language users do in non-test situations.

The strategy of elimination shares some features with the strategy of guessing. The basis for elimination as well as the way a test-taker makes use of the strategy seems to be the decisive factor. While for the introspective responses where guessing was said to be used the success rate was under 40 %, for elimination it is above 50 %, suggesting that it is, taken as a whole, a more reliable strategy than the different forms of guessing. It may also imply that it is a strategy that is more often combined with a more confident and covering comprehension of the spoken text.

Parallel to the cases with the strategy of guessing, the reasons for applying elimination varies – there is "high-level" elimination employed mostly by more successful test-takers, in 70 % of the cases leading to a correct choice. Test-takers have, on the basis of a fairly covering comprehension of the spoken text, compared their mental representation of the text with each of the options, eliminating the incorrect options. On the other hand, there are test-takers who use clues outside the text to rule out improbable options – especially in cases where there are comprehension problems. Sometimes the reasons for elimination seem to reflect the risk of construct-irrelevant variance, where the strategy is used in an "invalid" way. The problem with opaque options seems, as far as this strategy is concerned, to surface among the more serious threats to the quality of some of the items.

At best, elimination, inherent in the context of a MC task, is a process of comparing this alternative against that, on the basis of an interpretation of the text, an individual's background knowledge and some general cognitive factor – may it be called task-solving sensibility, as an element also related to target language use situations (cf. Haladyna 2004: 61).

An important point to make here is the necessity of being aware of the way we define "guessing" and "elimination". This conceptual question has to do both with what test-takers, as writers of the introspective responses, mean by "guessing" and "elimination", and what we as researchers decide to interpret as evidences of these strategies. As has been discussed in the context of the analysis related to these seventeen test items, these strategies are heterogeneous as what comes to their cognitive level, the ways of their employment as well as the consequences of their use.

The different introspective responses reflect different stages or levels in the listening test-taking process (See Figure 14 in Appendix 3). The results taken together suggest that there are elements in the process identified by the introspective responses that are common to both language test and non-test situations. Nevertheless, there is also clear proof of elements that cannot be claimed to be found in a target language use situation. There are construct-irrelevant factors as a function of at least two variables: the individual item and the individual test-taker. These two variables will be treated under the two following research questions.

10.2 How does the nature of the individual MC item influence the employed processes and strategies?

First of all, it is clear from the analysis that the nature of the individual MC item does influence the processes and strategies that the test-takers make use of. The one and same test-taker rarely gives proof of the same processes or the same strategies throughout the seventeen listening comprehension items. The nature of the individual item influences the processes and strategies activated with that particular item. What Rupp et al. (2006) has found in his study in the context of a MC test of reading comprehension has clear parallels to the context of listening comprehension. Rupp et al. (2006: 468) points out that it has to be shown empirically what type and level of comprehension a MC item assesses, since characteristics of the questions and the text interact with characteristics of the test-takers to induce response processes that are mediated by prior experience with such tests. Based on the introspective responses, in the following I will describe and discuss possible situations and circumstances evidenced as a reflexion of the nature of the items.

The factors that affect the nature of the studied items are many and varying, just like the elements that any piece of linguistic sample consists of are of varying character (as described in Table 9; Rupp et al. 2001). Every item is in some way different, which is usually expected when the objective is to cover as large a part of a construct as possible by a test that by necessity includes only a representative sample of all potential tasks within that construct: the items are made with the intention of testing different aspects of the same defined construct. The quality of the present items is expected to be relatively high, due to the fact that the items are created and used in a high-stakes examination context (see the construction process described in chapter 4.2).

By an analysis of the contents of the items (the spoken text passages and questions related to these), we have expectations on the functioning of the items. Some potentially difficult or problematic features are easy to foresee, some are not. There are textual factors that have an influence on the difficulty of an item – quantitative factors like text length and placement of necessary information, or qualitative factors like the characteristics of the vocabulary, or factors expressed by means of the concept of cognitive load referring to the informational structure of the text (cf. Rupp et al. 2001; Brown 1995). In the case of spoken language input there are naturally many traits related to the phonology, the pace, the clarity of pronunciation and the intonation or different *sandhi*-phenomena that are crucial factors with an influence on the difficulty also in the case of spoken French, as the language has many traits different from those found in the first languages of the test-takers (Finnish or Swedish).

As was established under the first research question, there are factors other than the spoken text that influence the processing. A test question can be asked in many ways, through many different formats. For the current MC questions, the fundamental task is, as was maintained, to understand the spoken text input and decide which of three alternative propositions corresponds to the text contents, or answers the question in the stem. The main difficulty of the task is not necessarily in understanding and interpreting the spoken text, but in the interaction of the text and the question (cf. Rupp et al. 2006; Yi'an 1998). One interactional feature adding to the cognitive load of the test-taker and the difficulty of the items is in the situation where the order of the questions does not follow the order of the information in the spoken text, or where as many as three questions are to be answered within the same time slot. Even a relatively difficult text (for example a relatively long text with abstract concepts) can yield relatively easy items, depending on what the question is. On the other hand, the contrary can be true. A relatively easy text for test-takers on a specific target level can yield demanding test items, either by purpose, or unintentionally through a flaw in the item. If the informational contents of the options are very close to each other, for example, only some detail in the options may relate to the correct interpretation of the text, and this detail may be difficult to perceive (Cf. Rupp et al. 2006). Unintentional difficulties are created by using vocabulary or concepts in the written options that are unfamiliar to the test-takers on the target level for the test. This makes the task opaque, and leads to a potential situation where a test-taker who has understood the text contents still cannot prove his or her understanding by selecting the key option, in case if he or she is unable to compare his or her mental text representation with the meaning of all of the three options. This situation may cause construct-irrelevant variance in the test results, and constitutes a real threat to the validity and the reliability of the item and thus of the test.

The nature of the item influences the way the task is processed, as judged by the introspective responses given by the 218 test-takers in this analysis, and this contributes to the outcome for a particular test-taker on a particular item. The traits of a spoken text related to a particular item influences how this text is processed, what seems salient to a test-taker and how much he or she finally understands and retains of the text contents. If the test-taker has had the chance to read through the task, the question and options, in advance, he or she has got both an artificial purpose for listening, as well as some expectations on what the text to be listened to may contain. The condition is, however, that the task is clear and transparent. To this adds the aspect of time pressure usually present in the context of listening comprehension assessment, where the chance of listening to the spoken text as well as the time provided for responding to the task is controlled and limited.

If the task imposed by the item is clear and the spoken text is experienced as being on a convenient level for the target group for the test, the processes and strategies for an average test-taker are rather straightforward. The task indicates what the test-taker is supposed to focus on in the spoken text, and the test-taker attempts to understand and interpret the text in the light of the task as thoroughly as he or she can. Depending on the amount and quality of what the test-taker has understood of the text, he or she solves the task as best he or she can, possibly using some relevant background knowledge of the thematic field. To this is added the help of the strategy of elimination where the options are all compared against what is retained from the text and against each other, in order to find the best alternative. If most of the text content is understood, the key option can be detected without trouble. If the text is misunderstood, or if too narrow a fragment of the text contents is understood, a distractor is selected. This is what happens in the ideal and valid case.

There are other possible scenarios. At one particular item, the generally proficient test-taker may face excessive difficulties with the text if it is simply on a level that is too demanding for the target group for the test. The entire focus and energy is set on trying to solve the task on the basis of clues in the text, in the contents of the options and in some surface test-wise clues. If the "valid" clues don't seem to aid, the test-taker turns to some "invalid" ones – which can be labelled test-wiseness without a relation to the understanding of the spoken text. In the end the only solution may be to guess, the outcome being uncertain. The feeling of confusion experienced for a problematic item is likely to influence the processing of the following tasks as well.

The feeling of confusion is probably even stronger in cases where the test-taker has understood (the most of) the spoken text, made an interpretation and created a mental representation of the necessary information, but where the options are not understood. In fact, in the situation where the options are read through already before the first listening, to give a purpose for listening and help make expectations on what the spoken text will treat, the listening process itself may be disturbed by the situation where the options are not understood or are misunderstood (cf. Yi'an 1998; Rupp et al. 2006). Sometimes the spoken text may give hints on what an unknown piece of vocabulary in the options may refer to. In other cases, the test-taker may face a situation where the text is understood, but where he or she cannot prove his understanding by means of selecting the key option. Again, test-wiseness, or the strategy of elimination may save the situation, or then the test-taker has to be content with random guessing. Construct-irrelevant variance is likely to occur.

The consequence of the presence of opaque and unclear questions or options is likely to lead to high item measures. These items are likely to be difficult for the majority of the test-takers, and due to the fact that the test-takers have to rely on random guessing, the generally proficient test-takers may end up with selecting a distractor, while less proficient test-takers may happen to select the key, which leads to low discrimination figures. Obviously, an item of this kind does not measure anything – or measures very little – of what is included in the construct. An invalid and unreliable test may be the outcome.

On the other hand, in case the item is too easy, even if the text is on a suitable level, but if the distractors are not plausible enough to attract weaker test-takers who have not understood the spoken text correctly, the situation that emerges is one where any interpretation of the text leads to the selection of the key. Even test-takers who have not fully interpreted the main information in the text correctly can eliminate implausible distractors[93]. The presence of construct-irrelevant easiness is likely.

As a whole, based on the introspective responses, we have different potential scenarios that lead to the selection of a distractor or a key. These are presented in the following table (Table 36) and can be compared with the figure illustrating the listening test-taking process (Figure 14 in Appendix 3).

TABLE 36 Reasons for selecting the key or a distractor

Reasons for/ Ways of selecting **the key option**	**Reasons for/ Ways of selecting** **a distractor**
• Comprehension of the spoken text • Partial comprehension of the spoken text • Ruling-out based on information in the spoken text • Ruling-out based on background knowledge and inference • Informed guessing • Ruling-out based on test-wiseness or construct-irrelevant clues • Random guessing • OR a combination of these	• No understanding of the spoken text • Misunderstanding of the spoken text • Insufficient understanding of the spoken text • Misunderstanding of the stem or the options • No understanding of the stem or the options • "Falsely informed" guessing or elimination (for example over-reliance on background knowledge on the expense of the textual information) • Random guessing • OR a combination of these

Some of these reasons seem valid, and we can consider that the one or zero (or three and zero within the Finnish Matriculation Examination context) points on an item deserve to be earned. In other cases, the processes and strategies leading to the response does not seem to be valid for the outcome.

In terms of the validation procedures and using the notions of claims and counterclaims (as in Bachman 2004), the central claim would be that the performance on the present test items is affected primarily by the test-takers' listening comprehension ability. We have some counterclaims to make regarding the factors that affect the test scores. These are related to what I have discussed above about reasons for selecting the key option or a distractor. Luck, test-wiseness and factors like excessively heavy information load, difficult vocabulary in the question or implausible distractors are obstacles to a valid and reliable test.

It can be said that there is not much that can be done to change the situation, that the problems are inherent in case the MC format is used. However, judging by the analysis of the introspective responses, when the processes and strategies leading to a response selection and a possible success are tracked down by means of the method of introspection, it seems clear that there is indeed something that can be done (and this is discussed in chapter 12.1 below).

[93] See for example Gao & Rogers (2011) on the importance and difficulty of creating plausible distracters.

10.3 How do the test-takers' listening processes and strategies relate to their success in solving the listening comprehension items?

The second variable affecting the processes and strategies employed in the test-taking situation is the test-taker (the first being the characteristics of the item and the text: cf. Rupp et al. 2006). There are naturally differences in the individual ways of handling a situation where a task is to be solved – there are more and less confident test-takers who react differently to possible problems, more or less flexible test-takers who can or cannot adjust their processes according to new information, and test-takers who are more or less dependent on the familiarity of a test situation or a test format. All these characteristics are bound to influence the test-taking and problem-solving situation for each individual test-taker (See Rantanen 2003: 187 and the present chapter 1.7).

However, the focus in the current study has been to investigate how the test-takers' conscious or unconscious choices of processes and strategies relate to the results in the current pool of items. Based on the results of the analysis of the introspective responses, certain tendencies are emerging. While the less successful test-takers tend to understand less of the text or misinterpret the intended message – visible through the introspective responses of word-bound and nonsense responses – there are gradually more cases of partial comprehension among test-takers with "intermediate" results and clearly more frequent résumés of the main contents of the spoken text by the strongest test-takers (see Table 25 in Appendix 4). This has an effect on the possible strategies used. The highest proportion of guesses is clearly found among the weakest test-takers, gradually decreasing towards the stronger test-takers. Elimination, on the contrary, is used most frequently among the strongest test-takers, while it is used in a decreasing proportion towards the weaker test-takers. This seems rather as expected in terms of first of all what test-takers with different levels of success are expected to grasp of the text, and second, how the strategies are employed on the basis of that comprehension. An important point is the fact that, as was established by Yi'an (1998), the point of departure for the weaker test-taker faced with a MC item is less advantageous already due to characteristics of the pre-view format: they cannot profit from the conditions in the same manner as the more advanced test-takers, for whom the pre-viewing facilitate the processing.

The metacognitive responses are given in rather equal proportions through the success levels, since this is a heterogeneous category consisting of very different kinds of responses, ranging from comments on the difficulty of the task to explanations on what basis for inference has been used at solving specific items. One point to make here is the fact that these referred results consist of averages of all test-takers with certain total scores or person measures. There are differences between individuals with the same person measure and individual differences as a function of differences in the item. Second, the generalizations are made and the big picture is drawn by adding up the separate responses that are given within a group of several test-takers with higher or lower person measures. This is due to the fact that there are not many cases where the individual test-taker would have described his or her test-taking processes and strategies exhaustively. The researcher has to assume that different individual test-takers inform us about different phases in the test-taking procedure, and that adding introspective responses tells about the entire picture. Other advantages and drawbacks of the method of introspection are discussed in the following chapter.

11 DISCUSSIONS ON THE QUALITY OF THE CURRENT RESEARCH STUDY

Having used the method of short written introspection and analysed the results both in a pilot study and in this current research study, what becomes clear for the researcher is that there are advantages and disadvantages associated with this specific method and the way it has been applied in this particular study.

The particular form of introspection used for this study – short written introspection, or "write-down-thinking" – has not been applied, or at least referred to before in the study of listening comprehension. The task is made as straightforward as possible for the test-takers and subjects as they are simply asked to state the reason for selecting a particular option on each item. This has two major benefits. First of all, the format allows for large groups of test-takers to take part, which gives the researcher the possibility of combining the qualitative information obtained by means of the verbal protocols with quantitative information on the items and the test-takers. Second, the task is made easy, which implies that it does not demand such an effort that the focus is taken completely away from the test-taking. In this manner, the task is likely to lead to fewer cases of abandon of the task and of empty responses.

The limited time and space for the introspective task play down the task further, and even if the responses are shorter and perhaps less "profound" compared with open oral verbal reports, they represent a reflection of what is at the top of the minds of the test-takers in the heat of the moment of solving the task determined by an item.

I will approach the methodological discussion by discussing the limitations of the method of introspection pointed out by the experts in the field. First of all, it is clear that not all cognitive processing in the test-taking situation is available for introspection – there are many tasks and processes taking place simultaneously, some of them being automatic and unconscious. For stronger test-takers, some processes may be so automatic, that they are not reflected on at all on a conscious level (see Nagle & Sanders 1986; Phakiti 2003). As is established by Brown (1995) and Yepes (2001) – it is mainly the problems and obstacles in the situation that attract the test-takers' attention and that make more controlled and conscious processing necessary. The lexicon is probably the most tangible element in the linguistic processing (cf. Buck 1990) and the part that is most easily also introspected on and written down. What the introspective responses did not manage to cover, were the many other linguistic aspects of the listening process proper. Before the pilot study, I would have expected to find as justifications for the selection of a particular option more cases of direct quotation or mishearing, or misinterpretations of the spoken text, and which would have been given in the target language. This might have provided more information on the different phenomenon in the process of listening to and interpreting a spoken text, ranging from phonological details to discourse patterns at large. However, due to various reasons, this was not the case.

As a first reason, it appears that listeners forget the exact wording of a spoken message quite quickly, in order to rationalize the processing, and only form a mental representation of what they have grasped of the message. This representation seems to, on the basis of the text-based introspective responses in this study, be stored in the listeners' mother tongue, the processing language, probably due to less cognitive load compared with a processing solely in the foreign target language. Second, we may speculate that in the case if the test-takers were to react orally to the text, by means of the think-aloud procedure, they may retain and express more elements directly from the spoken target language text, for example point to a word, or a sound sequence that is interpreted as a word or a phrase, that he or she has or has not understood. In oral verbal protocols, no spelling issues would be at stake, but only the phonological representation of a word or a sentence (fragment). Third, the fact that relatively little of the original text is reflected on, may also tell about the nature of the test-taking process, where, especially in a MC test, the focus is, possibly by necessity, set on the question and the options. This indicates an obvious challenge related to the testing of listening comprehension: how create test questions that do not interfere negatively with the understanding and interpretation of the spoken message, the core of the listening comprehension construct? It has to be pointed out that here the focus on the question was moreover emphasized by the research procedure, where the test-takers were to read through the questions already before the first listening. In the original test procedure, the question was studied only before the second chance to listen. However, as a compensating circumstance for the present analysis, in categorising the different introspective responses, priority was given to the textual elements that the test-takers reflected on. The advantage of having large groups of participants in the study is that the different ways of processing and reacting to the task at hand form a patchwork that gives an idea of the listening comprehension test-taking process in its entirety.

It is evident that the gathering of data during the test-taking event itself may interfere with the studied process (Cf. Alderson et al. 1995; Green 1997; Cohen 1998a). Even if this is inevitable, the statistical measure of a chi square test between the test results in the original administration of the test and the current test is conducted in order to give an idea of the comparability of the two test events. According to the results, in most of the seventeen items, there is no significant difference between the proportions of option choices for these items (See Table 16). It is mainly for two items that the proportions of option choices differ to a larger extent, due to circumstances discussed in chapter 5.2.2.

As far as the problematic issue related to the possibility of considering the current research study as a validation study is concerned, the first point relates to the use of only a part of the original test. If a validation study proper was the goal, the entire original test should be used. Otherwise, there may be some important parts of the test lost that would contribute to building a more thorough content coverage. Leaving out one part of the test makes the test an invalid and also unreliable validation study. Also, it is clear that seventeen MC items are too few to give a covering representation of a construct and too few to build a reliable estimate of the test-takers' level of skill on. As an afterthought, it would have been profitable not to change the four MC questions into open ended tasks (in the second part of the current "test" event), but instead to keep their original format. The total of 21 MC items would have allowed a more reliable idea of the test-takers' ability and would have given more data to analyse in order to cover an even larger part of the pattern of processing and strategy use. Another option would have been to include the open-ended items in the research study as a different (but unfortunately still an insufficient) dimension of the ability measure. Added to this, more external and background information on the participants and their results in other tests would be needed to be able to correlate the results of the test-takers with an external independent measure.

The number of test-takers constitutes a problem in a study that combines qualitative and quantitative data and methods. Even if the number of test-takers was sufficient for the qualitative analysis, the statistical analysis would have gained from a larger number of participants. On the other hand, with a restricted number of test-takers, more information from each could have been collected. This is a difficult compromise, but it is assumed that the analyses taken together complete each other in a fruitful manner.

Taking all these circumstances into account, it is clear that the approach - even if aiming at analysing the validity of the use of separate MC items - is taken from a restricted angle in the sense that it is the quality of the separate items taken as representing parts of a test of listening comprehension that is at stake.

The subjects and the data for the study are important factors related to the validity and the reliability of the results obtained from the study (see chapter 5.2). It is inevitable that the results are influenced by the restricted number of items selected for the study and by the students taking part. Nevertheless, in this case, the items represent a part of a test actually used for the purpose of assessing the listening comprehension ability of a target group with the same background as the group of test-takers used in the study. This also means that the items have already undergone important validation procedures at the stage of being created, evaluated and considered by the French division and the foreign language section of the Matriculation Examination board. The test circumstances are thus as closely as possible identical in the two situations, even if there are affecting factors that need to be taken into account in the study.

In considering the validity of this study from the point of view of changes in the test administration procedure, two issues need to be pointed out. First, a test situation set for research purposes where some kind of qualitative data is aimed at as a result can hardly be identical to the real test situation. There are always influencing factors that are not present in the test situation. This is in fact analogical to the relationship between a TLU situation and a test situation: a test is an "unnatural" constructed situation as much as a research study is. For the test, the targeted object is the ability; for the research study, the targeted object consists of the test or the items measuring that ability. The important issue to consider is whether the research circumstances altered the processes and strategies to such a degree, that completely different processes and strategies would be present if the test items were given in a "normal" test situation. I have judged that this is not the case, partly on the basis of the comparison of results to the items used in the original test and for the current research study (reported in chapter 5.2.2), partly on the basis of the nature of the introspective responses.

Second, there are always practical constraints influencing a research study, as there are practical constraints influencing a test situation. If it had been possible to use for instance 90 minutes instead of 45 minutes for the test, I could have included a more complete and even more purposeful test. On the other hand, this might have increased the fatigue factor towards the end of the test, influencing the results and outcome of the study. If it would have been possible to implement from a practical point of view, I could have covered and tested a much larger group of the future examinees with the intention to take part in that year's test of French in the Matriculation Exams (a total of approximately 3000 people). If given unlimited time to spend with the participants, I would have included a more covering oral retrospective interview with at least a limited group of participants.

As far as problems that may affect the validity of the verbal reports are concerned, it is clear that incomplete, distorted or extraneous reporting (Cf. Green 1997) is likely to occur. However, the method of short, written introspection with a large number of test-takers, allows, at least to some extent, the leveling out of these factors, so that some generalizations and tendencies can well be extracted from the results.

What is necessary to take into account is that the data provided by introspection demands some subjective interpretation and judgement by the researcher (Cf. Ross 1997). From the point of view of the validity it is important that the categorisation of the responses reflect the reported processes; the coding further need to be reliable. In the current study, the consistency of coding is verified by means of using two other independent coders for all the responses given by 226 participants to one of the items according to the coding instructions. In 71 % of the cases there was total agreement between all three coders, and in 90 % of the cases there was agreement reached between the researcher and at least one of the other coders (See chapter 9.2).

However, the researcher cannot overlook the challenges of the interpretation and categorisation of the introspective responses. The responses were asked for in an open format, the only practical restriction being the time and space available for the test-takers to respond. There are overlapping responses, and sometimes it was felt to be a question of taste or practicality to decide which category the responses would belong to: some responses were placed from one category to another as the analysis proceeded. The hierarchy of categories turned out to be necessary. The focus set on the questions – due to the test procedure – was partly compensated by the fact that the elements reflecting parts of the spoken text were judged higher in the hierarchy of categorisation.

In the closer investigation of the strategies of guessing and elimination, to the "pure" cases were added those including some element from another category. This was necessary for the sake of reaching a clearer understanding of these two strategies. For a follow-up study, I will have to reconsider and refine the principles of categorisation.

As a defence and in order to justify the open format, however, there were actually large similarities between the individual and independent responses, large enough to be used to generalize the results. Also, as the study combines quantity with quality, the introspective responses are described and analysed from several angles, and the presented examples serve as indications of how exactly the responses were worded. The objective of a deeper meaning of the quantitative information is thus reached by going into qualitative details.

If we consider the study from the point of view of alternative ways of collecting verbal protocol data, in comparison with an open think-aloud format, it is likely that not as abundant information is obtained by means of this method of written introspection as could be obtained by using spoken verbal protocols. However, it has been the attempt of the researcher to prove that this current format does provide more information than what would be expected. The advantages of the short written introspection is that the data is more easily handled, and thus can well be combined with a quantitative analysis, since large groups of test-takers can be tested. This serves to provide as complete a picture as possible of the processes and strategies and of the way these relate to item qualities and test-taker success.

Some researchers have preferred to use questionnaires where the processes and strategies are explicitly named for the test-taker to select from. These named strategies are naturally theoretically and empirically founded, and obviously result in statistically more manageable data. However, the drawback on served propositions is that they may make undue assumptions on the test-taker, who may want to respond according to the researcher's expectations. The closed format does not give the possibility for new, unexpected and individual responses to arise (See Cohen 1998a; Goh 2000). In a future study, the possibilities of combining a closed and an open format should be explored.

Another source of validity evidence would have been obtained by collecting more detailed information on the test constructors views and expectations on the functioning of each of the seventeen MC items, in the light of the results from the original test administration. They might have raised other important points of view and questions than those the researcher can think of.

In conclusion, the advantage of combining several types of data analysis is that there is an “internal” validity and reliability check, where results or tendencies obtained from one type of analysis are verified against the other data provided by the study. As has been shown, in some cases the functioning of an item can seem acceptable according to the quantitative data, but the qualitative data can reveal unexpected problems. Even if there are restrictive factors that have to be taken into account as far as the method of short written introspection and the limited number of items is concerned, the method has served its purpose of providing new, important information on the processes and strategies applied in the situation of solving MC items assessing listening comprehension. The results seem to be transferable both to testing contexts and to pedagogical situations.

12 IMPLICATIONS OF THE RESULTS OF THE CURRENT STUDY

There are potential implications of the results of the current study for the context of developing tests and writing listening comprehension test items as well as for the context of learning and teaching the skill of listening comprehension. This reflects another quality of the present research study: its usefulness and applicability to practical situations. In this chapter I will draw on the results of the present study and reflect on issues related to test development and pedagogical considerations.

12.1 Implications for the development of listening comprehension tests and item writing

For the test development context, there are a few issues and principles that seem to surface on the basis of the results in the study. However, as a general consideration, in a context where a test-taker's L2 ability as a whole is targeted, the basis for the development of a separate test of listening comprehension should build on the principle that there is a need for a separate test of listening comprehension. In many contexts, it may be profitable to integrate the assessment of listening comprehension with the assessment of (an)other skill(s), like speaking or writing for example. This is dependent on the construct and based on the ability level of the test-takers. If a separate test of listening is considered important in a particular context, it includes the assumption that a part of the L2 ability cannot be assessed by any other means than (or is best assessed) by a separate test. The listening comprehension construct thus includes features that cannot be reached by only the oral element, operationalized in a test of speaking, or by means of the element of comprehension, operationalized in a test of reading comprehension. The consequences of this assumption include the idea that the test of listening comprehension needs to include elements that belong to the listening comprehension construct, as limited as possible but at the same time as covering as possible. The objective is to avoid construct-underrepresentation and construct-irrelevant variance. For this purpose, the construct needs to be well defined in the test framework. However, this was not the case at the time of the original administration of the test of listening in spring 2002.

Next, the question arises what the test assessing the defined construct should be like. There are several task parameters, including the type of text and the type of questions asked (see Bejar et al. 2000). The parameters are interrelated – sometimes one type of text calls for one type of question format – and there are different kinds of practical restrictions influencing the decisions. In the current large scale assessment context, it seems profitable to combine the MC test format with open-ended short-answer questions, in order to capture a larger part of the listening comprehension construct and in order to avoid too large a method effect. Compared with the MC format, the short-answer questions allow the use of different types of texts as their bases leading to a different type of processing and the possibility of rewarding for partial comprehension. In the light of the introspective responses, this seems pertinent, as there appears to be an important difference between the situation where a part of the spoken text is understood, and the situations where either all or nothing is understood.

When the validity of the interpretation of MC test scores is the focus, the avoidance of the unreliable effects of random guessing should be one of the objectives. This can be reached by, first of all, using spoken texts that are within the scope of comprehension and interest of the targeted test-takers. The vocabulary and the informational cognitive load should not exceed what is expected of an average, or a slightly stronger test-taker. This has to be determined on the basis of the construct and possibly the curriculum that lies as a basis for the construct. It is essential to be aware of what is realistically expected to be mastered by the targeted group of test-takers. As can be concluded by the introspective responses to some of the analyzed items in the present study, an impossible text incites much random guessing, which distorts the results in many ways. Usually the original texts used for a listening test need to be rewritten with the focus set on the scope of comprehension for the target level of the test-takers. Apart from the need to include as genuine texts as possible, the selection of texts is limited by the informational contents of the text: some texts seem suitable for using for the development of MC questions while others are not: it has to be possible to find a clear key option and efficient distractors. The task for the test constructors is far from easy.

Second, the question (the stem and the options) need to be clear, transparent and preferably easily scannable. It has become very clear from the introspective responses that the question has a decisive effect on the rest of the processing. If the effect is great for the situation of a MC test of reading comprehension (as established by Rupp et al. 2006), it is even greater for a MC test of listening comprehension (Yi'an 1998), where the move between the text and the options is not possible, but where mnemonic restrictions and anxiety play an even more important part. The question gives an artificial purpose for listening. The test-takers need to know what is asked for, otherwise there is no purpose for listening and the entire process becomes distorted

For the item to be a proper MC task, and not several T/F tasks, the question should be asked in a straightforward way and not in a fuzzy and vague manner so that the test-taker knows what to focus on. Naturally, sometimes the question can concern the entire text passage where the options consist of alternative summaries of the main message in the passage. However, the options need to be real alternative responses, without the possibility of being simultaneously true.

The options not only have to be clear, but also plausible (see Gao & Rogers 2011). This is necessary in order to avoid construct-irrelevant test-wiseness affecting the test outcome in cases where the spoken text has not been understood, but where implausible options can be eliminated by means of "invalid reasons". It is, however, often demanding for the constructors to create plausible distractors that are sufficiently short, without vocabulary that is too demanding for the target level, but that manages to summarize the text contents, or possible misinterpretations of the test contents. It is often a game of paraphrases and synonyms.

Third, related to the two first issues on the text and the task, pre-testing is always to be recommended, since it is often hard to predict how the target group of test-takers will handle specific test items (Haladyna 2004). In the light of the current study, it can be recommended that quantitative results of the pretesting be completed with qualitative results, for example with the use some type of introspection, think-aloud or verbal protocol procedures that give information on each item separately. The challenge lies in avoiding too unwieldy procedures, but in developing ways of getting as much information as possible out of as little an effort as possible.

A question is whether it would be possible to use the method of short written introspection for validation purposes in the context of high-stakes and large-scale test development[94]. Judging by the results of this study, the introspective responses provide valuable information on the functioning of the items. The question is how large a group of test-takers would be needed in order to be able to gather reliable information on an item. Using 218 test-takers seems sufficient for the need of knowing how test-takers within the targeted group of potential test-takers use processes and strategies when faced with a particular item. This provides information on how an item functions, whether there are possible flaws or details to be improved. The more sources of information on the validity of an item, the better, and a triangulation with the analysis of the test contents combined with IRT (or some other type of quantitative item analysis) and some form of verbal protocol analysis seems like the optimal way of reaching a test that functions in a valid manner. But the question remains whether for example 50 test-takers' introspective responses would be sufficient for obtaining the needed information. Perhaps this situation with limited numbers of test-takers would be useful in an actual test validation context, at least in case of using sub-

[94] Issue also raised by Haladyna (2004: 197)

jects with very similar background and characteristics as the target population for the test. For reliable statistical analyses, larger groups would be needed.

However, as there are practical restrictions to the possibilities of using time, money and effort for the validation procedures, added to issues of secrecy, test developers might need to be content with using information from research where the method has been used, which provides some useful findings that are generalizable to their current testing context. The higher the stakes are, the more important the validation procedures. The more effort is put into the test and item development, the more valid and reliable results will be obtained. However, also for the learning context and assessment in class, it is important to use assessment practices that give a valid and reliable picture of the learners' skill. This has consequences on many levels for the individual who needs a true idea of his or her strengths and weaknesses.

12.2 Pedagogical implications

As far as the pedagogical implications and applications of the results of this study are concerned, what seems alarming is the gap between what the nature of the test (as it was conceived in 2002) and what a TLU context (and the construct as it is presents itself through the National Curriculum) would seem to suggest as being useful listening comprehension practice in class. The study of this test leads to the conclusion that practice in test-taking and techniques in solving MC items are important from the point of view of a successful test outcome. The teaching should cover strategies for monitoring one's performance and staying focused, as well as compensating tactics when the comprehension fails or is not sufficient. Informed guessing and elimination are essential strategies for solving MC test tasks, so these should be practiced in class. All these strategies are essential ingredients in the real-life language use context as well, but with a slightly different focus: the language itself more than the assessment instrument.

The starting point for learning L2 listening on the target level of the current test should be to learn to understand the language related to immediate needs, more concrete and communicative contexts than what the items that are included in the studied test would suggest. The language functions reflected through the present items do not seem particularly authentic or genuine, or seem to cover only a part of the construct. The National Curriculum stands in the context of the Matriculation Examination as an implicit framework enlightening the particular construct of listening comprehension that the test items should have as a target. "Everyday language" as well as communication strategies are focused as important concepts. This leads to the question what everyday spoken language is, and what features in this language are important. There are certainly domains and functions of different kinds, but what seems to be missing in the operationalization of the construct behind this test of listening comprehension is the characteristics of colloquial and unplanned discourse (described for example by Rost (2002: 123-124) and Buck (2001: 112). After all, this type of spoken language forms an essential part of what should be practiced in class in order for a learner to be able to cope with the L2 in a TLU situation. Therefore, there cannot be any doubts about the fact that this type of input should be recommended for class (or for an independent learner) on the present target level. The type of tasks combined with this input should be as varied as possible, in order to have the effect of interactiveness and in order to develop purposeful strategies.

In order to manage the listening comprehension test situation as it presented itself in the 2002 version, the type of language and tasks that are recommended for practice are different, because the needed processes and strategies are of a different kind. The importance of the familiarity with possible test formats should not be forgotten – it is more likely that a test-taker can make use of all his or her linguistic processing capacity if he or she does not have to worry about some unexpected features of the test. A teacher thus has to balance the limited hours spent in class on helping the learners develop partly different types of skills needed for listening in the test situation and listening in a potential real-life context. With the present items in mind and knowing by experience, it is possible that test-takers can get relatively high scores in a MC test of listening, but have severe difficulties in understanding the spoken French in real life situations.

As far as practicing or assessing the skill of listening comprehension in class is concerned, various types of diagnostic assessment seems more worth an effort than summative MC tests as tools for knowing where the strengths and weaknesses of a learner lie. The learner-strategy approach (see chapter 1.2) and the combination of introspection with Dynamic Assessment (see chapter 1.1) seem interesting in this context. The fact that the individual level and progress of a particular learner are taken into account, and that interaction and negotiation are key elements seem to make this approach very useful for the context of learning correctly targeted strategies that can be applied both in TLU and test situations. This naturally demands a lot of training and experience from the teacher. Peer cooperation is another recommendable way of learning useful strategies from each other, as the hours spent in class are scarce and as the teacher needs to divide the limited time between all learners, often in large groups. Added to that, introspection seems to be a helpful tool in learning purposeful monitoring and self-evaluation in the test situation as well as in a non-test situation. In the comments provided at the end of the test form, many test-takers told that they appreciated the task of explicitly reflecting on the reasons behind their option selection, considered by many to be a useful exercise.

Teachers are usually faced with the need to give a grade to their learners, and this on the basis of some tests in class. This grade is usually a general grade on the L2, without specifications on the abilities in the separate skills. It may thus be meaningful to assess integrated skills, combining listening comprehension with for example speaking or writing. Relevant tasks seem rather easy to find, if only the criteria for giving grades are clearly and explicitly determined[95]. Based on the results of the present study, creating MC questions for assessing separate language skills in the classroom context does not seem worth the effort, considering how demanding it is to develop valid and reliable MC items.

[95] See Takala 1998 on the comparison of traditional and alternative assessment methods.

13 SUGGESTIONS FOR FURTHER RESEARCH

There seem to be several openings related to the context of this study where further research would be needed or would indeed be interesting. One study that is already piloted[96] concerns distractor plausibility and the use of open-ended questions to compare open responses with the suggested MC response options. The target would be to find reasonable key options and plausible distractors for the MC format. This type of study could be developed further, with the exploration of different types of questions – focusing on detailed or global information for example. This would be a possible parallel approach that could be combined with short written introspection, in order to reach a very covering idea of how a spoken text is processed. The obtained information could be used in helping learners develop purposeful processes and strategies as well as in developing test items.

Another approach would be to investigate the validity and reliability of MC test items assessing listening comprehension, where partial scores would be given for option choices evidencing partial comprehension of the spoken text. The demands of the type of text and the options allowing for partial scores would be different than for a traditional MC test, and it would be necessary to look into differences in the test-takers' processing and use of strategies as well.

If we part from the results obtained in the present study, there seems to be two major ways of proceeding. The first concerns the deepening of the study, implying that more profound information is obtained on the processing of each of the test items. This could be done by means of using oral verbal protocols and a more limited number of test-takers on different sublevels. This would on one hand provide more information on the entire processing and employment of strategies for solving a test task by one individual test-taker. On the other, with several items, and combined with more background information on the test-taker this would provide test-taker profiles as a function of some characteristic of the test-taker for instance. Tests of listening comprehension in several

[96] Anckar: *Is Your Test Question Related to My Answer? Exploring the Relationship Between Test-takers' Interpretations of Spoken Text and the MC test Format.* Paper presented at the fifth annual EALTA Conference 2008 in Athens, Greece.

languages for one individual test-taker could be compared: are the employed processes and strategies transferable across languages? With the present context with (at least) two very different L1, a further dimension would be the exploration of differences in the processes and strategies as a function of different L1.

The second way is to create questionnaires with lists of processes and strategies for example on the basis of the results obtained so far. This would be a quantitatively larger study, where the test-takers would select their responses separately for each item, thus providing information on each. The lists should include the dimensions of text processing (from the level of the word to the "necessary information"), strategies (guessing and elimination) and other reactions and reflections. The number of test-takers should be in hundreds, and the questionnaires would be filled in during the test event.

As far as the strategy of guessing in listening comprehension is concerned, there are several further dimensions to be explored. What this strategy implies for the cognitive processing for a learner, whether in a test situation or another language use situation, would be an interesting area to explore more profoundly. The approach could be either an assessment perspective or a more pedagogical one and cover also what I have labelled elimination and inferencing in the present study.

Differences in processes as a function of different test formats would be another object of study. The processing activated for MC-, T/F, multiple true-false (MTF) and open-ended formats could be compared across comparable groups of test-takers. There still remains to be found the "best approach" to assess listening comprehension ability, and proceeding through investigating the processing will be fruitful. Another approach related to the development of the methodology of short written introspection would be to try it out with tests of other language skills. Imaginable tests would at least be MC-, T/F, and MTF tests of reading comprehension, as well as cloze tests assessing grammar and vocabulary, both with open cloze, MC and MTF format.[97] The processes and strategies are expected to be very different as a function of the skill to be tested and the format of the test.

Related to the present high-stakes context, the Finnish Matriculation Examination, a contradictory situation exists because of the demands of on the one hand, the examination system and, on the other hand, the foreign language construct and curriculum. Today, at the construction stage, the foreign language test items are related to the criterions and descriptions in the *CEFR* in order to create more purposeful and comparable references to international language competence levels. However, the test-takers' results in the foreign language tests are not related to the criteria, but to the norms of the score system that are dependent on the results of other test-takers on the same test items. There is consequently a gap between the test-taker's real competence level and

[97] Research for example on the MTF format in assessing reading comprehension and vocabulary knowledge is reported in Dudley, A. Multiple dichotomous-scored items in second language testing: investigating the multiple true–false item type under norm-referenced conditions. *Language Testing*, 2006, Vol. 23 Issue 2, p198-228

the scores given by the assessment machinery, creating an invalid system that should be changed. This is a very important subject for further studies, with potential important consequences for the assessment system. Research into different aspects of this problem should combine statistical methods with qualitative methods, in order to reach convincing results to use for proving the need for and the possibility of change.

Another area of current interest is the possibility of integrating the assessment of oral production and oral comprehension. The need of assessing learners' L2 speaking skills within the context of the Matriculation Examination is long since admitted[98], but the practical ways of operationalizing this need are yet to be found. The possibility of using computer-based testing in this context is an important research subject.

All this leads to the conclusion that even if the current study has succeeded in completing the information on the processes and strategies employed by learners in a test of listening comprehension of a foreign language on the characteristics of the MC test items as well as on the usefulness of the introspection method for item validation purposes, there are still vast territories left to be explored in order to reach the potential point where the processes and the products meet.

[98] See for example Saleva (1997).

TIIVISTELMÄ

Vieraan kielen kuullun ymmärtämistaidon mittaaminen monivalintatehtävien avulla: prosesseja ja tuloksia

Tässä väitöskirjatutkimuksessa tarkastellaan ranskan kuullunymmärtämiskokeeseen osallistuvien kokelaiden tulosten takana piileviä ymmärtämisprosesseja ja strategioita. Tutkimuksessa yhdistyvät seuraavat näkökulmat: kuullunymmärtämisprosessin moniulotteisuus ja Buck:in (2001) kuvaama kuullunymmärtämisen viitekehys, kuullunymmärtämistaidon arvioinnissa käytettävien monivalintatehtävien tiedostetut ja piilevät ongelmat, sekä osioiden validius ja validiuden varmistaminen. Näistä taustatekijöistä nousee kolme tutkimuskysymystä: 1) Mitä prosesseja ja strategioita kokelaat käyttävät ratkoessaan seitsemäntoista ranskan kuullunymmärtämisen monivalintaosiota? 2) Miten monivalintaosion ominaisuudet vaikuttavat käytettyihin prosesseihin ja strategioihin? 3) Miten käytetyt prosessit ja strategiat liittyvät kokelaiden koesuoritukseen ja tuloksiin?

Keskeisenä menetelmänä tutkimuksessa on introspektio, jossa koehenkilöt testitilanteessa ilmaisevat ajatuksiaan lyhyesti ja kirjallisesti (*short written introspection*). Tutkimuksessa käytetyt osiot ovat lähtöisin kevään 2002 ylioppilaskirjoitusten lyhyen ranskan kuullunymmärtämiskokeesta. Tutkimukseen osallistui 218 ranskaa opiskelevaa abiturienttia 22 suomalaisesta lukiosta, äidinkielenään suomi tai ruotsi. Opiskelijat vastasivat seitsemääntoista kuullunymmärtämisen monivalintaosioon samalla perustellen valintojaan osioittain. Näistä perusteluista koostuu introspektioaineisto, joka valottaa opiskelijoiden käyttämiä prosesseja ja strategioita.

Introspektioaineiston mielekkyyden varmistamiseksi osiot analysoitiin kvalitatiivisin ja kvantitatiivisin (Raschin IRT) menetelmin. Näin tutkittiin, mitkä tekijät ranskankielisessä kuullussa tekstissä, monivalintakysymyksissä ja vastausvaihtoehdoissa vaikuttanevat osion vaikeustasoon. Lisäksi selvitettiin, että osiot keskimäärin vastaavat kokelaiden osaamistasoa.

Introspektiovastauksista käy ilmi sekä kuullun tekstin prosessointiin liittyvät asiat, strategioiden käyttö että kokelaiden tunneperäiset reaktiot koetilanteessa. Introspektiovastaukset osoittavat, että esimerkiksi arvaaminen ja poissulkemistaktiikka eivät ole ”huonoja” tai ”epärehellisiä” taitoja, vaan ovat luontevia vastaamis- ja ymmärtämisstrategioita. Tällöin arvaus ei siis perustu satunnaisuuteen vaan tietoon (*informed guessing*). Poissulkemistaktiikassa puolestaan vastausvaihtoehtoja suljetaan kuultavan tekstin ja kuulijan tiedon perusteella eikä niinkään kuultavaan liittymättömien vihjeiden perusteella. Sekä arvaaminen että poissulkeminen voivat olla hyödyllisiä strategioita myös autenttisissa kielenkäyttötilanteissa.

Sekä kokelaiden että osioiden ominaisuudet vaikuttavat siihen, mitkä prosessit ja strategiat kulloinkin aktivoituvat (vrt. Rantanen 2003; Rupp et al. 2006; Yi’an 1998). Osioihin useimmin oikein vastanneet kokelaat osaavat muodostaa

suhteellisen kattavan mielikuvan kuullusta tekstistä ja vertaavat sitä jokaiseen vastausvaihtoehtoon. Näin he voivat sulkea pois harhauttajat ja valita oikean vastausvaihtoehdon. Osioihin väärin vastanneet kokelaat puolestaan keskittyvät kuulemiinsa yksittäisiin sanoihin joita he yhdistävät vastausvaihtoehdoissa esiintyviin sanoihin. He myös käyttävät poissulkemisstrategiaa ja arvaamista kompensoivina strategioina silloin, kun mitkään tekstistä ymmärretyt sanat eivät johda oikean vastauksen tielle. Jos kuulija ei ymmärrä kuulemaansa ollenkaan, hän tukeutuu arvaamiseen tai vihjeisiin, joita ei esiinny kuullussa ollenkaan. Nämä koetilanteessa ilmenevät eri ”käyttäytymismallit” voivat muuttua yksittäisen osion ominaisuuksien perusteella.

Tulosten mukaan kognitiivisesti vaativa teksti (vrt. *cognitive load*, Brown 1995), epäselvät kysymykset tai vaihtoehdot (Rupp et al. 2006, Yi’an 1998) tai liian lähellä oikeaa vastausta olevat tai epätodennäköiset harhauttajat johtavat ”vääristyviin” prosesseihin ja strategioihin. Jos vastausvaihtoehtoja ei ymmärretä, tästä voi seurata, että tekstin hyvinkin ymmärtäneet kokelaat eivät pysty osoittamaan ymmärtämistään, vaan heidän on käytettävä strategioita, jotka ovat erittäin kaukana siitä, mikä olisi luonnollista oikeassa kielenkäyttötilanteessa. Tällaisissa tapauksissa arvaamisesta ja poissulkemisesta tulee satunnaista, koska näiden strategioiden käyttö ei voi pohjautua tekstin ymmärtämisen ja vaihtoehtojen vertailuun. Toisaalta, jos osiot ovat liian helppoja epätodennäköisten harhauttajien takia (jotka siis eivät houkuttele edes heikkoja ymmärtäjiä), myös kokelaat, jotka eivät ymmärrä tekstiä, voivat vastata näihin oikein. Näin osioiden tehtävänä oleva heikkojen ja vahvojen ymmärtäjien erottelu ei toteudu.

Introspektiota käytettiin tässä tutkimuksessa siten, että kokelaita ei pyydetty kuvaamaan kaikkea, mitä he ajattelevat, vaan yksinkertaisesti perustelemaan itsenäisesti, lyhyesti ja kirjallisesti vastausvalintaansa. Näin toteutettuna introspektiotehtävä on epäsuora ja kokelaille suhteellisen helppo. Kokeen voi suorittaa suurikin ryhmä kerrallaan, ja vastaukset ovat lyhyitä ja suhteellisen helposti käsiteltäviä. Näin laatu ja määrä täydentävät toisiaan hedelmällisellä tavalla. Tämä tutkimus osoitti, että menetelmä toimii hyvin tutkimuksessa, jossa yhdistetään kvalitatiivista ja kvantitatiivista tietoa.

Tutkimus valottaa kokelaiden kuullunymmärtämistilanteessa käyttämiä prosesseja ja tätä kautta osioiden toimivuutta ja niitä tekijöitä, jotka vaikuttavat osioiden laatuun ja kokelaiden koetuloksiin. Tätä tietoa voidaan hyödyntää niin puheen ymmärtämisen kokeiden laatimisessa kuin puheen ymmärtämisen opettamisessa.

REFERENCES

Alderson, J.C. 1990. Testing Reading Comprehension Skills. Part Two. Getting students to talk about taking a reading test (pilot study). *Journal of Reading in a Foreign Language, 7,* p. 465-503.

Alderson, J.C. 2000. *Assessing Reading*. Cambridge: Cambridge University Press.

Alderson, J.C. C. Clapham & D. Wall. 1995. *Language Test Construction and Evaluation*. Cambridge Language Teaching Library. Cambridge: Cambridge University Press.

Alderson, J.C. 2004. In Cheng, Watanabe and Curtis (Eds.) *Washback in Language Testing*. (Foreword: ix)

Allan, A.I.C.G. 1992. *EFL reading comprehension test validation: investigating aspects of process approaches*. Unpublished PhD thesis, Lancaster University.

Allan, A.I.C.G. 1995. Begging the questionnaire: instrument effect on readers' responses to a self-report checklist. *Language Testing* 12 (2), 133-156.

AERA (American Educational Research Association), Americal Psychological Association & National Council on Measurement in Education. 1999. *Standards for educational and psychological testing.* Washington, DC: Author.

Anastasi, A. 1986. Evolving Concepts of Test Validation. *Annual Review of Psychology* 37, p.1-15

Anckar, J.M. 2003. *Validiteten och användbarheten i flervalstest i hörförståelse i franska som främmande språk.* Postgraduate Thesis. Åbo Akademi University.

Anderson, N.J, L. Bachman, K. Perkins & A. Cohen . 1991. An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources. *Language Testing,* Vol. 8, No. 1, 41-66.

Anderson, J.R. 1985. *Cognitive Psychology and Its Implications*. New York: Freeman.

Anderson, A. & Lynch, T. 1988. *Listening*. New York: Oxford University Press.

Bachman, L.F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L.F. 2004. *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

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

Bachman, L.F. & A.D. Cohen (eds). 1998. *Interfaces between second language acquisition and language testing research.* Cambridge: CUP

Bailey, K. M. 1998. *Learning About Language Assessment*. Cambridge, MA: ITP/Heinle& Heinle.

Baker, R. 1997. *Classical test theory and item response theory in test analysis.* Lancaster University. Centre for research in Language Education.

Bejar, I., D. Douglas, J.Jamieson, S. Nissan, & J. Turner. 2000. *TOEFL Listening Framework: A Working Paper.* TOEFL Monograph Series MS 19.

Bialystok, E. 1990. *Communication strategies: a psychological analysis of second-language use*. Oxford: Basil Blackwell.

Biber, D. 1995. *Dimensions of Register Variation. A Cross-Linguistic Comparison.* Cambridge: CUP.

Biber, D. 1997. Lexical bundles: What the grammar books won't tell you. In: *Perspectives on spoken and written discourse.* Colloquium conducted at the meeting of Teachers of English to Speakers of Other Languages, Orlando, FL.

Borsboom, D., G.J. Mellenbergh & J. van Heerden. 2004. The Concept of Validity. *Psychological Review*, Vol 111(4), p. 1061-1071.

Brindley, G. 1998. Assessing Listening Abilities. *Annual Review of Applied Linguistics.* Vol. 18, p. 171-91.

Brown, G. 1995. Dimensions of difficulty in listening comprehension. In: D.J. Mendelsohn, J. Rubin, (eds). *A Guide for the Teaching of Second Language Listening.* Carlsbad: Dominie Press, Inc.

Brown, G., A.H. Anderson, N. Shadbolt & T. Lynch. 1985. *Listening Comprehension.* Project JHH/190/1. Edinbourgh: Scottish Education Department.

Buck, G. 1990. *The Testing of Second Language Listening Comprehension.* Unpublished Doctoral Thesis. Lancaster University.

Buck, G. 1991. The testing of listening comprehension: an intospective study. Language Testing, vol. 8, No.1, p.67-91.

Buck, G. 1994. The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing* 11 (2), p. 145–70.

Buck, G. 2001. *Assessing Listening.* Cambridge: Cambridge University Press.

Bygate, M. 1987. *Speaking.* New York: Oxford University Press.

Canale, M. and Swain, M. 1980. Theoretical bases of communicative approaches to second lanaguage teaching and testing. *Applied Linguistics 1*, p. 1-47.

Carrell, P. 1988. Some Causes of Text -Boundness and Schema Interference in ESL Reading . Carrell, P., J.Devine and D.Eskey (red.*) Interactive Approaches to Second Language Reading*). New York : Cambridge University Press

Cavalcanti, M. C. 1987. *Investigating FL reading performance through pause protocols*. In: C. Faerch & G. Kasper (eds). *Introspection in Second Language Research*. Clevedon: Multilingual Matters.

Cervantes & Grainer, G. 1992. The effects of syntactic simplification and repetition on listening comprehension. *TESOL Quarterly, 26, p.* 767-770.

Chamot, A.U. 1995 Learning Strategies and Listening Comprehension. In: D.J. Mendelsohn and J.Rubin (eds). *A Guide for the Teaching of Second Language Listening*. Carlsbad: Dominie Press.

Chang, L., Y. Watanabe & A. Curtis (eds). 2004. *Washback in Language Testing.* Research Contexts and Methods. Lawrence Erlbaum and Associates.

Chaudron, C. 1983. Simplification of Input: Topic and Reinstatements and Their Effects on L2 Learners' Recognition and Recall. *TESOL Quarterly* 17: p. 437-58

Chaudron C. 1995. Academic listening. In: D. Mendehlsohn and J. Rubin (eds). *A guide for the teaching of second language listening.* (p.74-96) San Diego, CA. Dominie press, Inc

Chaudron, C. & J. Richards. 1986. The Effect of Discourse Markers on the Comprehension of Lectures. *Applied Linguistics* 1986 7(2): p.113-127. OUP.

Cheng, L., Yoshinori Watanabe & A. Curtis. (eds). 2004. *Washback in language testing. Research contents and methods.* Hillsdale, NeJ: Lawrence Erlbaum Associates,

Chiang, C. S. & Dunkel, P. 1992. The Effect of Speech Modification, Prior Knowledge and Listening Proficiency on EFL Lecture Learning. *TESOL Quarterly* 26: p. 345-74.

Cohen, A. 1998a. *Strategies in Learning and Using a Second Language*. Harlow: Longman.

Cohen, A. 1998b. Strategies and processes in test-taking and SLA. In: L. Bachman & A. Cohen (eds). *Interfaces between Second Language Acquisition and Language Testing Research.* CUP.

Cornaire, C. 1998. *La compréhension orale*. Paris: CLE Internationale

Derwing, T.M. 1989. Information type and its relation to non-native speaker comprehension. *Language Learning*, 39, 2, p. 157-172.

Derwing, T.M. 1996. Elaborative Detail. Help or Hindrance to the NNS Listener? *SSLA*; 18: p. 283-297.

Dickinson, L. 1987. *Self-instruction in language learning.* Cambridge: Cambridge University Press.

Dijk, T. A. van & W. Kintch. 1983. *Strategies of Discourse comprehension*. New York: Academic Press.

Dirven R.& J. Oakshott-Taylor. 1984. Listening comprehension (Part I). *Language Teaching* 17: 326-43.

Downing S.M & T. M. Haladyna (eds). 2006. *Handbook of Test Development.* Mahwah, NJ: Erlbaum

Ebel , R.L. & D.A. Frisbie. 1991. *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall

Ellis, R. 1994. *The study of Second Language Acquisition.* Oxford: Oxford University Press

Emslie, J.R & G.R Emslie. 2005. *Improving Classroom Multiple-Choice Tests: A Worked Example Using Statistical Criteria*. Department of psychology: Ryerson University Toronto.

Ericsson, K.A & H.A. Simon. 1987. Verbal Reports on Thinking. In: C. Faerch & G. Kasper (eds). *Introspection in Second Language Research.* Clevedon: Multilingual Matters.

Educational Testing Service. 2003. *ETS Standards for fairness and quality*. Princeton, NJ.

Faerch, C. & G. Kasper . 1987. From Product to Process – Introspective methods in Second Language Research. In: C. Faerch & G. Kasper. (eds) *Introspection in Second Language Research.* Multilingual Matters : Clevedon.Firth, A. & Wagner, J. (1997). In an article in the Modern Language Journal (1997, 81, p. 285-300, republished in *The Modern Language Journal*, 91, 2007)

Flowerdew, J. & L. Miller. 2005. *Second Language Listening. Theory and practice*. New York: Cambridge University Press.

Freedle R. & Kostin, I. 1999. Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing,* 16 (1), p. 2-32

Fulcher, N. G. & Davidson, F. 2007. *Language testing and assessment: an advanced resource book*. London: Routledge.

Gao, L. & W.T. Rogers. 2011. Use of tree-based regression in the analyses of L2 reading test items. *Language Testing* 28 (1): p. 77-104.

Geranpayeh, A. & L. Taylor. 2008. Examining Listening: developments and issues in assessing second language listening, *Cambridge ESOL Research Notes* Issue 32 May.

Glisan E.W. 1985. The effect of word order on Listening Comprehension and Pattern retention: An experiment in Spanish as a Foreign Language. *Language Learning 38:* p. 443-72.

Goh, C.C. M. 2000. A cognitive perspective on language learners' listening comprehension problems. *System*. 28, no. 1, p.55-75. Oxford ; New York : Pergamon Press.

Goh, C. C. M. 2002. Exploring listening comprehension tactics and their interaction patterns. *System*. 30, no. 2, (2002): p.185- 206. Oxford ; New York : Pergamon Press.

Green, A. 1998. *Verbal Protocol Analysis in language testing research: A handbook*. Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Greene, J. 1986. *Language Understanding: A Cognitive Approach*. Milton Keynes: Open University Press.

Haastrup, K. 1987. Using Thinking Aloud and Retrospection to Uncover Lerners' Lexical Inferencing Procedures. In: C. Faerch & G. Kasper (eds). *Introspection in Second Language Research*. Multilingual Matters : Clevedon

Haladyna, T.M. 1994. *Developing and Validating Multiple-Choice Test Items*. Mahwah, NJ: Erlbaum.

Haladyna, T.M. 2004. *Developing and Validating Multiple-Choice Test Items*. (3rd ed.) Mahwah, NJ: Erlbaum.

Haladyna, T.M., S.M. Downing & M.C Rodriguez . 2002**.** A Review of Multiple - Choice Item-Writing Guidelines for Classroom Assessment**.** *Applied Measurement in Education,* vol 15, Nr 3 p. 309-34.

Hambleton, R.K. & L. N. Murrey. 1983. Some goodness of fit investigations for item response models. In: Hambleton, R.K.(ed.) *Applications of Item Response Theory.* Vancouver: Educational Research Insitute of British Columbia.

Hansen C. & Jensen, C. 1994. Evaluation of lecture comprehension. *Academic Listening,* p. 241-268). New York: Cambridge University Press

Henrichsen L. 1984. Sandhi-variation: A filter of input for learners of ESL. *Language Learning, 34,* p. 103-126.

Huhta, A. & M. Tarnanen. 2009. *Assessment practices in the Finnish comprehensive school – what is the students' role***?** (Huhta, A. & M. Tarnanen (accepted). Assessment practices in the Finnish comprehensive school - what is the students' role in them? LED 2007 Conference Proceedings.)

Jamieson, J., S. Jones, I. Kirsch, P. Mosenthal, & C. Taylor. 2000. *TOEFL 2000 Framework.* A Working paper. Monograph No. MS-16

de Jong, J.H.A.L & C.A.W. Glas. 1987. Validation of listening comprehension tests using item response theory. Language Testing, Vol. 4 no. 2, p. 170-194

Kane, M. T. 1992. An Argument-Based Approach to Validity. *Psychological Bulletin.* Vol 112, No.3: 527-535. American Psychological Association.

Kauppinen, M., J. Saario, A. Huhta, A. Keränen, M.-R. Luukka, S. Pöyhönen, P. Taalas & M. Tarnanen. 2008. Kielten oppikirjat tekstimaailmaan ja – *toimintaan sosiaalistajina.* [Language study books as a way of becoming socialized into textual world and language use]. Jyväskylä; AFinla. AFinla Yearbook No 66.

Kelch, K. 1985. Modified Input as an Aid to Comprehension. *Studies in Second Language Acquisition* 7: p. 81-89.

Kelly, P. 1991. Lexical ignorance: the main obstacle to listening comprehension with advanced foreign language learners. *International review of Applied Linguistics.*29 (2) p. 122-4.

Kreiter, C.D. & Frisbie, D.A. 1989. Effectiveness of Multiple True-False Items. *Applied Measurement in Education*; Vol. 2 Issue 3, p. 207-217

Kyriacou, C. , N. Benmansour & G. Low. 1996. Pupil Learning Styles and Foreign Language Learning. *Language Learning Journal*, vol 13, issue 1, p. 22-24.

Larsen-Freeman, D. 2007. Reflecting on the Cognitive-Social Debate in Second Language Acquisition. *The Modern Language Journal*, vol. 91, Issue s1, p. 773-787.

Lhote E. 1995. *Enseigner l'oral en interaction.* Paris: Hachette.

Linn, R.L. & M. D. Miller. 2005. *Measurement and Assessment in Teaching* (9th ed.) Columbus, OH: Pearson Education.

Livingston, S.A. 2006. Item Analysis. In: Downing S.M & T. M. Haladyna (eds). *Handbook of Test Development.* (p. 421-443). Mahwah, NJ: Erlbaum

Luoma, S. 2001. What does your test measure? Construct definition in language test development and validation. PhD thesis. Jyväskylä University.

Lynch, T. 1998. Theoretical perspectives on listening. *Annual Review of Applied Linguistics.* 18, p. 3-91.

Malmberg, B. 1976. *Phonétique française.* Malmö: Liber.

McNamara, T. F. 1991. Test dimensionality: IRT analysis of an ESP listening test. *Language Testing*, 8: p. 139-159.

Mendelsohn, D.J. & J. Rubin (eds). 1995. *A Guide for the Teaching of Second Language Listening.* San Diego, CA: Dominie Press

Messick, S. 1989. Validity. In: R.L. Linn (ed.) *Educational Measurement* (3rd ed., p. 13-103). New York: Americal Council on Education, Macmillan Publishing Company.

Messick, S. 1994. *Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning*. Research Report RR-94-45.

Moss, H. E., & Gaskell, M. G. 1999. Lexical semantic processing during speech. In Garrod, S. & Pickering, M. (eds). *Language Processing.* Hove: Psychology Press.

Nagle, S. & S. Sanders. 1986. Comprehension theory and second language pedagogy. *TESOL Quarterly,* 20.1.

Nakatari, Y. 2006. Developing an Oral Communication Strategy Inventory. *Modern language Journal* 90, vol. 2, p. 151-168.

Nevo, N. 1989. Test-Taking Strategies on a Multiple-Choice Test of Reading Comprehension. *Language Testing*, Vol 6, Nr 2 p.199-215

Nissan S., F. De Vincenzi. & K.L Tang. 1996. *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension.* (TOEFL Research Report No. 51). Princeton, NJ: Educational Testing Service.

Noblitt, J.S. 1995. Cognitive Approaches to Listening Comprehension. *Dimension*, p. 1-11. *(Proceedings of the 1995 Joint Conference of the Southern Conference on Language Teaching and the South Carolina Foreign Language Teachers' Association.)*

Norris, J.M., J. D. Brown, T. Hudson & J. Yoshioka. 1998. *Designing Second Language Performance Assessments.* University of Hawaii at Manoa. National Foreign Language Resource center.

Oller, J.W. 1979. *Language Tests at School*. London: Longman.

O'Malley, J.M., Chamot, A.U. & Küpper, L. 1989. Listening comprehension strategies in second language acquisition. *Applied Linguistics*, 10(4), p. 418-437.

O'Malley, J. M., & Chamot, A. U. 1990. *Learning strategies in second language acquisition.* Cambridge, England: Cambridge University Press

Oxford, R. 1990. *Language Learning Stretegies. What every Teacher should know.* NY: Harper & Row/ Newbury House.

Oxford, R. 2003. Language learning styles and strategies: Concepts and relationships. *International Review of Applied Linguistics* 41(4), p. 271–278.

Oxford, R., Cho, Y., Leung, S. & Kim, H-J. 2004. Effect of the presence and difficulty of task on strategy use: An exploratory study. *International Review of Applied Linguistics* 42 (1), p. 1-47

Phakiti, A. 2003. A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading comprehension test performance. *Language Testing,* 20(1), p. 26-56.

Pica T., R.Young & C. Doughty. 1987. The Impact of Interaction and Comprehension.*TESOL Quarterly* 21, p. 737-58.

Poehner, M.E. & J.P. Lantoft. 2005. Dynamic assessment in the language classroom. *Language Teaching Research*, Vol. 9, No. 3, p. 233-265.

Powers, D. 1985. *A survey of academic demands related to listening skills.* (TOEFL Research Report 20). Princeton, NJ: Educational Testing Service.

Purpura, J. 1999. *Strategy use and second language test performance: A structural equation modeling approach.* Cambridge: Cambridge University Press.

Rantanen, P. 2003. *Enemmän vähemmällä. Monivalintatehtävien mittaustarkkuuden nostaminen*. Kasvatusalan tutkimuksia. Turku: Suomen Kasvatustieteellinen seura.

Rodriguez M.C. 2005. Three options Are Optimal for Multiple-Choice Items: A meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice,* vol. 24, Nr 2 p. 3-13.

Ross, S. 1997. An introspective analysis of listener inferencing on a second language listening test. In: Kasper, G. & E. Kellerman (eds) *Communication strategies: Psycholinguistic and Sociolinguistic Perspectives*. (p. 216-237). Addison Wesley: Longman

Rost, M. 1994. *Introducing Listening.* Penguin English.

Rost, M. 1999. Developing Listening Tasks for Language Learning. *Odense Working Papers in Linguistics*. University of Odense.

Rost, M. 2002. *Teaching and Researching Listening.* London: Longman.

Rost, M. & S. Ross 1991. Learner Use of Strategies in Interaction: Typology and Teachability. *Language Learning*, Volume 41, Issue 2. p. 235-73.

Rubin J. 1994. A review of second language listening comprehension research. *Modern Language Journal,* 78, p. 199-221.

Rumelhart, D.E. & Ortony, A. 1977. The representation of knowledge in memory. Anderson, R.C., Spirox, R.J & Montague, W.E. (red.) *Schooling and the acquisition of knowledge*. Hillsdale, NeJ: Lawrence Erlbaum Associates.

Rupp, A.A., T. Ferne & H. Choi. 2006. How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language Testing*, Vol. 23, No. 4, p. 441-474.

Rupp, A.A., P. Garcia & J. Jamieson . 2001. Combining Multiple Regression and CART to Understand Difficulty in Second Language Reading and Listening Comprehension Test Items. *International Journal of Testing*, 1(3&4), 2001, p.195-6.

Schank, R. 1975. The structures of episodes in memory. In: Bobrow, D. & A. Collins. (Eds.) *Representations and Understanding* .

Shannon, C.E. & W Weaver. 1949. *The mathematical theory of communication,* Urbana: University of Illinois Press

Shohamy E. & O. Inbar. 1991. Validation of Listening Comprehension Tests: The Effect of text and question Type. *Language Testing* 8, p. 23-40.

Stemmer, B. 1991. *What's on a C-test-taker's mind*. *Mental Process in C-test taking.* Bochum: Brockmeyer.

Takala, S. 1998. Language Testing. Recent Developments and persistent Dilemmas. In: *Puolin ja toisin: Suomalais-virolaista kielentutkimusta. AFinLAn vuosikirja 1998* (*On Both Sides: Finnish-Estonian Research on Language.* AFinLA Yearbook 1998). Paper presented at "Linguistics in Estonia and Finland: Crossing the Gulf" Symposium (Tallinn, Estonia, November 14-15, 1997).

Tannen, D. 1982. The oral/literate continuum of discourse. Tannen, D. (ed.) *Spoken and written language: exploring orality and literacy.* Norwood, NJ: Ablex Publishing Co

Tarnanen, M., Huhta, A. & Pohjala, K. 2007. Mitä on osaaminen? Kielitaidon arviointi vastaajana. In: Pöyhönen, S & Luukka M-R. (Eds.) *Kohti tulevaisuuden kielikoulutusta. Kielikoulutuspoliittisen projektin loppuraportti*.

Tarone, E. 2007. Sociolinguistic Approaches to Second Language Acquisition Research 1997-2007. *The Modern Language Journal*, 91, p. 837-848.

Tegelberg, E. 1995. *Franskt uttal i teori och praktik*. Göteborg: Akademiförlaget.

Thompson I. 1995. Assessment of second/foreign language listening comprehension. In: D. Mendelsohn & J.Rubin (eds) *A Guide for the teaching of second language listening* (p. 31-58). San Diego, CA: Dominie Press, Inc

Tsui, A.B. & J. Fullilove. 1998. Bottom-Up or Top-Down Processing as a Discriminator of L2 Listening Performance. *Applied Linguistics* 19/4, p. 432-451.

Turner, K. 1995. *Listening in a foreign language – a skill we take for granted?* Pathfinder 26. London: CILT.

Ur, P. 1984. *Teaching listening comprehension.* New York: Cambridge University Press.

Vandergrift, L. 1992. The comprehension strategies of second language (French) listeners. Unpublished doctoral dissertation, University of Alberta, Canada.

Vandergrift, L. 1996. The listening comprehension strategies of core French high school students. *Canadian modern language review* 52, p.22, 200-223. University of Toronto Press.

Vandergrift, L. 1997. The comprehension strategies of second language (French) listeners: A descriptive study. *Foreign Language Annals* 30, 3, p.387-409.

Vandergrift, L. 1998. Successful and less successful listeners in French: What are the strategy differences? *The French Review,* 71 (3), p. 370-95.

Vandergrift, L. 1999. Facilitating second language listening comprehension: acquiring succcessful strategies. *ELT Journal* 1999 53(3), p.168-176. Oxford University Press.

Vandergrift, L. 2003. Orchestrating Strategy Use: Toward a Model of the Skilled Second Language Listener. *Language Learning*, Vol 53, Issue 3, p. 463 – 496.

Vandergrift, L. 2005. Relationships among Motivation Orientations, Metacognitive Awareness and Proficiency in L2 Listening *Applied Linguistics,* 2005 26(1), p. 70-89.

Wagner, M.J. 2006. Utilizing the visual channel : an investigation of the use of video texts on tests of second language listening ability. Thesis. Teachers College, Columbia University.

Weir, C. J. 2005. *Language Testing and Validation: an evidence-based approach.* Basingstoke : Palgrave Macmillan.

Yanagawa, K. & A. Green. 2008. To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System* 36, p.107-122.

Yepes, J. 2001. *Using analysis of retrospective interviews following a TOEFL listening task to refine a model of L2 listening comprehension*. Paper presented at the 2001 AAAL Conference. St Louis, Missouri.

Yi'an, W. 1998. What do tests of listening comprehension test? – A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, Vol. 15, nr 1: p.21-44.

Young, M.Y.C. 1997. A Serial Order of Listening Comprehension Strategies Used by Advanced ESL Learners in Hong Kong. *Asian Journal of English Language Teaching,* vol. 7, p. 35-53. CUHK English Language Teaching Unit.

The Common European Framework of Reference: Learning, Teaching, Assessment. (2001). Council of Europe Council for Cultural Co-operation. Education Committee. Modern Languages Division (Strasbourg), Cambridge: CUP.

APPENDIX 1

The first sheet of the test paper. Instructions originally in the test-taker' mother tongue, Finnish or Swedish, here translated by the researcher.

Test of listening comprehension of French, 15.1 2004.	
I	
You are going to hear the following texts one at a time divided in passages. You will listen to each text twice. When you listen to the text the first time, answer the questions during the (70 sec) pause: at this stage you are to circle the option that seems correct to you. After having listened a second time, justify your selection: briefly write down on what basis you have arrived at your choice – or indicate if you made a guess. After the second listening you can also change your choice of an option: mark the new choice by an asterisk (*).	
	Justification:
1.De quel phénomène parle-t-on ici? a) La mort des petites boutiques b) L'attrait des logements au niveau de la rue c) Les prix élevés des rez-de-chaussée	
2.Comment explique-t-on ce phénomène? a) C' est uniquement une question d'argent b) Le centre-ville est devenu trop cher c) Les gens cherchent de nouveaux liens sociaux	
3.Comment Joël décrit-il sa vie? a) Il aime discuter avec les touristes b) Il s'occupe des chômeurs du quartier c) Il a envie de vivre dehors	
4.Quel est le comportement des gens avec lui? a) Ils le dérangent dans son travail b) Ils le prennent pour un photographe c) Ils le traitent comme un touriste	
5.Que raconte Martine? a) Les bruits de la rue ne la dérangent pas b) Elle vient d'avoir des problèmes de santé c) Elle veut repartir en province	
6.Que dit-elle de leur appartement? a) Il n'est pas trop cher b) Il leur évoque des souvenirs c) Il ne sera jamais prêt	

Text passage and MC items assessing listening comprehension of French as a L2, originally used in the Finnish Matriculation Examination in spring, 2002

<table>
<tr><td><i>1. De quel phénomène parle-t-on ici?
a) La mort des petites boutiques
b) L'attrait des logements au niveau de la rue
c) Les prix élevés des rez-de-chaussée</i></td><td><i>2. Comment explique-t-on ce phénomène?
a) C' est uniquement une question d'argent
b) Le centre-ville est devenu trop cher
c)Les gens cherchent de nouveaux liens sociaux</i></td></tr>
<tr><td colspan="2"><i>Aujourd'hui, à Paris, on habite dans des boutiques. C'est un phénomène recent et il s'explique par plusieurs raisons: Les petits commerces ferment. Le rez-de-chaussée ne fait plus peur. Les appartements ordinaires sont devenus trop chers. Les agences immobilières n'hésitent plus à vendre d'anciennes boutiques. La motivation économique est évidente, mais ce n'est pas la seule raison pour laquelle on s'intéresse à ce nouveau mode de vie. C'est un acte volontaire, social. C'est comme avoir une maison en ville, dans un immeuble. On cherche la vie communautaire, l'esprit de la campagne. Aujourd'hui on vit dans un esprit d'ouverture et de rencontres.</i></td></tr>
</table>

<table>
<tr><td><i>3.Comment Joël décrit-il sa vie?
a) Il aime discuter avec les touristes
b) Il s'occupe des chômeurs du quartier
c)Il a envie de vivre dehors</i></td><td><i>4. Quel est le comportement des gens avec lui?
a) Ils le dérangent dans son travail
b) Ils le prennent pour un photographe
c) Ils le traitent comme un touriste</i></td></tr>
<tr><td colspan="2"><i>Joël, 53 ans, est professeur de théâtre et vit dans un ancien magasin de fleurs, à Montmartre.
« Quand arrivent les beaux jours, c'est génial. Je sors ma table, je travaille sur le trottoir. Je suis entouré d'arbres et de plantes, je me crois à la plage ! L'inconvénient, c'est qu'il y a pas mal de gens qui ne travaillent pas dans le quartier et qui viennent me parler quand je travaille et que je veux être tranquille. Alors je suis désagréable. En général, ils comprennent. Ici, c'est très touristique et les gens me photographient comme une personnalité bizarre, ce qui m'ennuie. Mais quand je suis enfermé, je ne me sens pas à l'aise. »</i></td></tr>
</table>

<table>
<tr><td><i>5.Que raconte Martine?
d) Les bruits de la rue ne la dérangent pas
e) Elle vient d'avoir des problèmes de santé
f) Elle veut repartir en province</i></td><td><i>6.Que dit-elle de leur appartement?
d) Il n'est pas trop cher
e) Il leur évoque des souvenirs
f) Il ne sera jamais prêt</i></td></tr>
<tr><td colspan="2"><i>Olivier et Martine ont tous les deux 35 ans. Lui est directeur commercial, elle, consultante en informatique. Ils ont acheté un vieux laboratoire de prothèses dentaires dans le 17ème arrondissement.
« Nous sommes originaires du Midi avec mon mari et ici, ça nous rappelle un peu la vie de province. Deux vitrines donnent sur la rue, les autres sur une cour et un jardin…mais quand j'ai essayé de dormir dans une pièce du côté de la rue, c'était trop bruyant, entre les poubelles le matin et les rondes des agents de police. L'achat a bien sûr demandé de gros investissements – et pas seulement financiers. J'ai fini par me sentir au bout de mes forces, crevée ! J'ai perdu plusieurs kilos, mais le pire, c'est que mes nerfs ont complètement craqué… Mais maintenant, après les débuts difficiles, ça a l'air d'aller bien. Le plus dur est derrière nous. »</i></td></tr>
</table>

7. Qu'apprenons-nous sur la vie de Cyril et Cécile? *a) Ils laissent leurs amis profiter de leur espace* *b) Il y a un jardin d'enfants chez eux* *c) Ils sont propriétaires d'une galerie d'art*	*8. Qu'en est-il des cambriolages?* *a) On leur a déjà vole des tableaux* *b) La porte doit être toujours fermé* *c) Leurs parents surtout en ont peur*	*9. Quel inconvenient mentionne Cyril?* *a) La proximité de la rue, certains soirs* *b) Leur voisin jouant du piano la nuit* *c) Les passants entrant chez eux sans frapper*

Cécile, journaliste, et Cyril, éducateur, habitent dans un bar avec leurs deux petites filles. « Quand nous nous sommes installés ici, ce sont nos parents qui s’inquiétaient, à cause de cambriolages, et ils ne sont toujours pas tranquilles. Nous, ça ne nous a jamais fait peur. La porte est souvent ouverte. Il y a de la place, alors on a exposé des œuvres des amis peintres, des copains photographes. Quand un jardin d’enfants devait fermer, á cause de travaux, on l’a installé ici. Ça a duré deux mois et demi. Un inconvénient?Les soirs de fête, je dois surveiller la rue et regarder ce qui se passe, parce qu’il peut y avoir des bagarres. Ce qui est plutôt drôle, c’est le matin quand les gens, en rentrant chez eux, viennent demander un café! C’est sûrement à cause de l’enseigne « Bar de l’aventure ». Mais ça n’arrive pas souvent. Un soir, il s’est passé un truc génial. Un mec avait entendu le piano. Il est rentré chez nous et il s’est mis à jouer. C’était sympa. »

10. Il y a longtemps, l’appartement de Micheline était…

a) …un bureau de poste
b) …un restaurant
c) …un magasin d’alimentation

11. Quel souvenir d’enfance Stéphanie a-t-elle gardé?

A Elle jouait souvent dans la rue
B Elle n’avait pas d’amis
C Elle ne voyait jamais sa mère

Micheline, 57 ans, vit avec son mari et leur fille Stéphanie, 29 ans. Ils vivent depuis trente-deux ans dans cette ancienne épicerie, mais eux, ils n’ont jamais rien vendu. « J’ai voulu retrouver ma Normandie, l’esprit de la campagne. C’est une petite rue, tout le monde se connaît. J’ai un tableau avec pas mal de clés, je réceptionne des colis et du courrier pour les gens, je les aide à remplir leurs papiers. Les gens viennent me voir pour tout ! Le soir, quand on est á table il peut arriver que des gens passent la tête et demandent : Il est complet, votre restaurant ? »

Stéphanie raconte : « Moi j’avoue que, souvent, j’aurais bien aimé avoir ma mère pour moi seule. On lui demandait tout le temps quelque chose ! A part ça, j’ai vécu une enfance superbe. Je sortais ma couverture sur le trottoir et j’étalais mes jouets, mes poupées. J’ai été très entourée par les voisins et les commerçants et je n’ai jamais senti la solitude des enfants uniques. »

25. Qu’est-ce que le Rayon Vert a de spécial?
a) Il est connu pour la qualité de son café
b) Il accueille des artistes peu connus
c) Il s’adresse à des non-spécialistes

Les galeries d’art contemporaine font souvent un peu peur. On aimerait entrer, et pourtant, on n’ose pas. La galerie le Rayon Vert a décidé d’accueillir mieux les non-initiés. Avec ses ouvertures d’expositions café-croissant, acceuillant tous les passants et ses rencontres avec les artistes, le Rayon Vert a déjà réussi à attirer un nouveau type de publique.

26. De quoi s’agit-il?
a) D’une idée de Descartes
b) D’un découverte scientifique
c) D’un nouveau prix scientifique

Matématicien, philosophe, voyageur curieux, humaniste du 17ème siécle, René Descares est dèsormais associé à une récompense scientifique ambitieuse. Le prix européen qui porte son nom, veut figurer parmi les plus grands palmarès internationaux.

27. Quelle serait la réplique suivante?
a) « Ne quittez pas «
b) « Ne laissez pas »
c) « Ne passez pas »

- Crédit Lyonnais, bonjour.

- Oui, euh, bonjour. Je viens de récevoir une lettre de vous, et j’aimerais demander quelques renseignements supplémentaires. Pouvez-vous me passer la personne qui s’occupe des comptes privés, s’il vous plaît?

28. *Quelle serait la réplique suivante?*
a) « Non, je ne le savais pas. »
b) « Si, c’est obligatoire. »
c) « Non, c’est toi qui ne le sais pas. »

- *Dis, donc, il y en a du monde ici!*
- *Ça n’a rien d’étonnant. Tout le monde veut faire la même chose que nous, partir le plus vite possible pour profiter au maximum des vacances. Allez, dépêche-toi, on n’a qu’à monter là.*
- *Oh, j’espère qu’on trouvera des places assises. Je n’ai pas pensé à faire une reservation…*
- *Comment? Tu sais pourtant bien que c’est obliagatoire dans les TGV!*

29. *Quelle serait la réplique suivante?*
a) « Non, je n’aime pas l’entrecôte.»
b) « D’accord, je pense que ça ira. »
c) « Merci, je prends du filet. »

- *Bonjour, monsieur. Je voudrais un morceau de viande pour mon chat.*
- *Et qu’est-ce que vous voulez comme morceau?*
- *Du filet, naturellement!*
- *Il n’y a plus de filet, madame, prenez de l’entrecôte!*

30. *Quelle sérait la réplique suivante?*
a) « Ça aurait été chouette. »
b) « Je n’ai pas de vacances. »
c) « J’espère que ça marchera. »

- *Écoute, Fabienne, j’ai une idée. Cette année nous passons les vacances en Vendée. On a loué une grande maison au bord de la mer pour tout le mois d’août. Je me suis dit que tu pourrais venir nous rejoindre pour quelque jours.*
- *C’est une idée formidable! J’aime la mer. On s’amuserait ensemble. Mais avant de répondre, il faut que je parle avec ma famille et que je vois en quelle moment je peux prendre des vacances*

Text passage and MC items assessing listening comprehension of French as a L2, items originally used in the Finnish Matriculation Examination in spring, 2002, here translated to English by the researcher

1. What phenomenon is discussed here? a) The death of small shops b) The attraction of flats on the street level c) The high prices of the ground floors	2. How is this phenomenon explained? a) It is only a question of money b) The town centre has become too expensive c) People look for new social relationships

Today, in Paris, people live in shops. It is a recent phenomenon and is explained by several reasons. The small shops are closing down. The street-level is not frightening anymore. Ordinary appartments have become too expensive. The housing agencies don't hesitate to sell old shops anymore. The economic motivation is evident, but this is not the only reason why people are interested in this new way of living. It's a volontary act, a social one. It's like having a house in town, in a block of flats. People look for community life, the countryside spirit. Today people live in a spirit of openness and meetings.

3. How does Joël describe his life? a) He likes to discuss with the tourists b) He takes care of the unemployed in the neighbourhood c) He wants to live outdoors'	4. What is the behaviour of people with him? a) They disturb him in his work b) They take him for a photographer c) They treat him like a tourist'

Joël. 53 years old, is a drama teacher and lives in an old florist's shop in Montmartre. "When the beautiful days come, it's great. I take out my table, I work on the pavement. I'm surrounded by trees and green plants, as if I was on a beach! The inconveniancy is that there is quite a lot of people who don't work in the area, and who come talking to me when I work and would like to be left alone. Then I'm being impolite. Usually theyn understand. It's very touristy here and people take pictures of me as if I was a bizarre character, which irritates me. But when I'm locked indoors, I don't feel good".

5. What does Martine tell? a) The noises in the street do not disturb her' b) She has just had some health problems c) She wants to go back to the countryside	6. 'What does she say about their apartment? a) It is not too expensive b) It evokes some memories for them c) It will never be ready'

Olivier and Martine are both 35 years old. He is a manager, she is an consultant in IT. They've bought an old denture laboratory in the 17th arrondissement. "My husband and I we come from the south of France, so this reminds us a little of countryside life. Two windows face the street, the other ones the courtyard and a garden… but when I tried to sleep in a room on the street-side, it was too noisy, between the dustbins in the morning and the rounds by the police. The purchase naturally demanded big investments – and not only financial ones. In the end I felt completely exhausted, finished! I lost several kilos, but the worst was that by nerves broke down completely… But now, after the difficult start, all seems to be going well. The hardest times are behind us."

7. What do we get to know about the life of Cyril and Cécile? a) They let their friends use their space b) There is a kindergarten in their house c) They are the owners of an art gallery	8. How is it with the burglaries? a) Pictures have already been stolen from them b) The door always has to be closed c) Their parents are especially afraid	9. Which inconvenience does Cyril mention? a) The nearness of the street, certain nights b) Their neighbour playing the piano at night c) The passers-by who come in without knocking

Cécile, journalist, and Cyril, teacher, live in a bar with their two little daughters. “When we moved in, our parents were worried, because of burglaries, and they’re still not calm. We’ve never been afraid. The door is often open. There’s plenty of space, so we’ve exhibited pieces of art made by our painter and photographer friends. When a kindergarten had to be closed, due to some renovation, it was placed here. It took two and a half months. Some inconveniency? On party nights, I have to watch the street and see what happens, since there may be fights. What’s rather fun is when in the mornings when people are heading back home they drop in and ask for a coffee! It must be because of the sign “Adventure Bar”. But that doesn’t happen often. One night, a brillliant thing happened. A guy had heard the piano. He entered our house and started to play. That was nice.

10. A long time ago, Micheline’s apartment was…
a) …a postal office
b) …a restaurant
c) …a grocery shop

11. What childhood memory has Stephanie retained?
a) She often played in the street
b) She didn’t have any friends
c) She never saw her mother

Micheline, 57 years old, lives with her husband and her daughter Stéphanie, aged 29. They’ve lived for thirty-two years in this old grocery, but they’ve never sold anything. “I wanted to find my Normandy, the countryside spirit. It’s a small street, everyone knows eachother. I’ve got a board with quite a few keys, I take care of people’s packages and mail, I help them fill in papers. People come to see me for everything! In the evening, when we’re seated at the table it happens that people stick in their heads asking: Is it fully booked, your restaurant?”

Stéphanie tells: “I can admit that often I would have liked to have my mother all to myself. People were asking her something all the time! Apart from that, I’ve had a wonderful childhood. I took out my blanket on the pavement and spread out my toys and my soft toys. I was surrounded by neighbours and merchants and I’ve never felt the lonelyness of an only child.”

25. What is the speciality of le Rayon Vert?
a) It is famous for the quality of its coffee
b) It welcomes less known artists
c) It is intended for non-specialists

Modern art galleries are often a bit frightening. One would like to enter, but still one doesn’t dare. The le Rayon Vert- gallery has decided to be more welcoming towards outsiders. With their exhibitions “café-croissant’ that welcome all passers-by, and their meetings with the artists, le Rayon Vert has already manage to attract a new type of audience.’

26. What is this about?
a) One of Descartesis ideas
b) A scientific discovery
c) A new scientific price

Matematician, philosopher, a curious traveller, a humanist of the 17th century, René Descartes is from now on associated with an ambitious scientific award. The European price that carries his name, aims to be one of the greatest on the international top list.

27. Which would be the following line?
a) (Don’t leave →) “Hold on”
b) “Don’t leave”
c) “Don’t pass”

– Crédit Lyonnais, good morning!

- Yeah, good morning. I just received a letter from you, and I'd like to ask for some supplementary information. Could you connect me to the person that takes care of private accounts, please?

28. Which would be the following line?
a) "No, I didn't know that"
b) "Yes, indeed, it is compulsory"
c) "No, it's you who don't know"

- I say, it is crowded here!

- That's not surprising. Everyone wants to do the same as us: leave as early as possible to get the most out of the holidays. Come, on, hurry up, we'll just jump on here!
- O-oh, I hope we'll find seats for us. I didn't think about making a reservation…
What? But you do know it's compulsory on the TGV!

29. Which would be the following line?
a) "No, I don't like entrecote"
b) "Okey, I think it's fine"
c) "Thank you, I'll take the filet"

- Good afternoon, sir. I'd like a piece of meat for my cat.
- And what kind of meat do you want?
- Filet, naturally!
- There is no filet left, ma'm, have some entrecote instead!

30. Which would be the following line?
a) "That would have been nice"
b) "I don't have any vacation"
c) "I hope that it will work out"

- Listen, Fabienne, I've got an idea. This year we spend our holidays in Vendée. We've rented a big house by the sea for the entire month of August. I thoughtf that you could come and join us for a few days.
- That's a great idea! I love the sea. We'd have fun together. But before I give an answer, I have to talk to my family to see when I can take a holiday.

APPENDIX 2

Quantitative information on the current 17-item MC test of listening comprehension.

```
               <more>|<rare>
   80             .  +
   79            .#  +
   78                +
   77                +
   76                +
   75                +
   74                +
   73                +
   72             .  +
   71                +
   70                +
   69                +
   68               T+
   67             #  +
   66                +
   65                +
   64                +
   63          .###  +T item25
   62                +
   61                +
   60         #####  +
   59             . S+
   58                +  item27
   57         #####  +
   56                +S item6
   55                +
   54    .#########  +  item3   item4
   53                +  item10
   52                +
   51       #######  +  item29  item9
   50               M+M item1
   49   .##########  +  item5
   48                +  item8
   47                +  item2   item26  item7
   46    ##########  +
   45             .  +
   44                +S
   43        ######  +  item28
   42                +
   41        ###### S+  item30
   40                +
   39                +
   38                +
   37           ###  +T
   36                +  item11
   35                +
   34                +
   33          .###  +
               <less>|<frequ>
(. = 1-2 cases; # = 3 cases)
```

FIGURE 6 Person-item map showing the distribution of item difficulty and person ability on the same scale

TABLE 17 Proportion of selection of options for each item in the original and the current test administration. The key option for each item is underlined

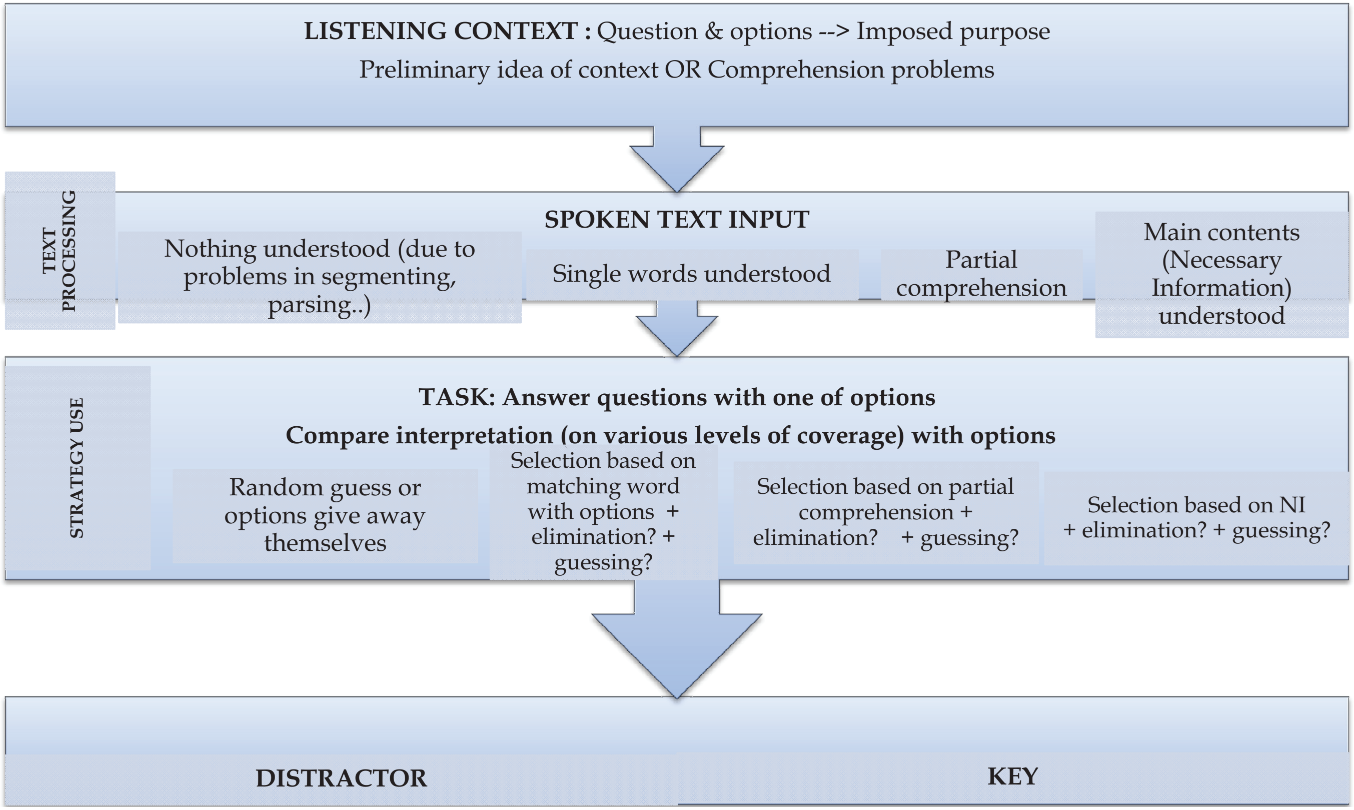
Item	Option	All Original 2002 Test-takers N: 3262 % selected	Current administration 2004-2005 Test-takers N: 218 % selected
1	a	32.5	39.9
	b	57.4	48.6
	c	10	11.5
2	a	11.1	17.4
	b	18.8	27.5
	c	70	55.
3	a	22.7	22.5
	b	37.5	35.8
	c	39.8	41.7
4	a	46	40.4
	b	37.5	39.9
	c	16.5	19.7
5	a	18.8	17.4
	b	56.7	51.4
	c	24.3	31.2
6	a	39.7	30.7
	b	35.4	37.2
	c	24.9	32.1
7	a	56.9	56
	b	16.3	20.2
	c	26.7	23.9
8	a	18.2	20.8
	b	28.3	25.9
	c	53.4	53.2
9	a	55.3	47.2
	b	13.3	15.6
	c	31.4	37.2
10	a	13.4	15.1
	b	44.9	42.2
	c	41.6	42.7
11	a	76.2	76.1
	b	15.3	11.5
	c	8.4	12.4
25	a	7.8	10.1
	b	63.5	66.5
	c	28.7	23.4
26	a	21.6	22.5
	b	18.1	20.6
	c	60.2	56.9
27	a	32.7	33
	b	32.6	27.1
	c	34.5	39.9
28	a	67.7	64.2
	b	22.5	25.2
	c	9.7	10.6
29	a	44	47.7
	b	50.6	47.7
	c	5.3	4.6
30	a	22.7	28.4
	b	3.1	4.1
	c	74	67.4

TABLE 20 Results of the Rasch analysis of the 17 listening comprehension test items

name	measure	count	score	s.e#	inmnsq	inmnzemp	outmnsq	outmzemp	ptmeasur	obs match	exp match	discrimn
item1	50,4	18	106	1,46	0,99	-0,10	0,97	-0,43	0,38	65	65	1,04
item2	47,4	18	120	1,47	0,94	-1,36	0,89	-1,42	0,43	70	65	1,28
item3	53,7	18	91	1,49	1,11	2,07	1,11	1,41	0,27	58	67	0,63
item4	54,4	18	88	1,49	0,99	-0,15	0,97	-0,36	0,39	66	67	1,04
item5	49,2	18	112	1,46	0,99	-0,11	1,00	-0,03	0,37	65	65	1,01
item6	55,9	18	81	1,52	1,02	0,38	0,99	-0,04	0,36	65	69	0,96
item7	47,0	18	122	1,47	1,05	0,94	1,03	0,38	0,32	63	65	0,83
item8	48,0	16	116	1,47	0,95	-1,17	0,92	-1,11	0,42	67	65	1,24
item9	51,1	18	103	1,47	1,06	1,32	1,11	1,54	0,30	61	65	0,72
item10	53,2	18	93	1,48	1,00	0,04	0,98	-0,24	0,38	66	66	1,01
item11	36,5	18	166	1,69	1,02	0,21	0,97	-0,16	0,31	76	77	0,99
item25	63,4	18	52	1,72	1,05	0,58	1,04	0,37	0,32	78	78	0,93
item26	46,6	18	124	1,47	0,97	-0,60	1,14	1,74	0,37	70	66	1,03
item27	58,3	18	71	1,56	1,06	0,86	1,11	1,12	0,31	70	72	0,85
item28	42,8	18	141	1,52	0,92	-1,50	0,86	-1,46	0,43	74	69	1,26
item29	50,9	18	104	1,47	0,95	-0,99	0,93	-0,93	0,42	69	65	1,20
item30	41,4	18	147	1,55	0,97	-0,52	0,92	-0,77	0,38	70	71	1,10

APPENDIX 3

FIGURE 14 Hypothetical optional processes and strategies at stake in a MC test-taking situation.


LISTENING CONTEXT : Question & options --> Imposed purpose
Preliminary idea of context OR Comprehension problems
TEXT PROCESSING
SPOKEN TEXT INPUT
Nothing understood (due to problems in segmenting, parsing..)
Single words understood
Partial comprehension
Main contents (Necessary Information) understood
STRATEGY USE
TASK: Answer questions with one of options
Compare interpretation (on various levels of coverage) with options
Random guess or options give away themselves
Selection based on matching word with options + elimination? + guessing?
Selection based on partial comprehension + elimination? + guessing?
Selection based on NI + elimination? + guessing?
DISTRACTOR
KEY

APPENDIX 4:

Quantitative information related to the introspective responses

TABLE 25 Correlation between estimated person measure and the number of the different introspective responses per test-taker (N=218)

	Estimated Person Measure
Empty Pearson Correlation	.013
Sig. (2-tailed)	.850
Guess Pearson Correlation	-.294**
Sig. (2-tailed)	.000
Metacognitive Pearson Correlation	-.005
Sig. (2-tailed)	.946
Nonsense Pearson Correlation	-.070
Sig. (2-tailed)	.303
Option-focused Pearson Correlation	.067
Sig. (2-tailed)	.327
Partial comp Pearson Correlation	.146*
Sig. (2-tailed)	.032
Résumé Pearson Correlation	.598**
Sig. (2-tailed)	.000
Word-bound Pearson Correlation	-.283**
Sig. (2-tailed)	.000
Vague Pearson Correlation	-.063
Sig. (2-tailed)	.356

**** = correlation is significant at the 0.02 level (2-tailed)**
*** = correlation is significant at the 0.05 level (2-tailed)**

TABLE 26 The different types of introspective responses correlated with the item measure

	Estimated Item Measure:
Empty Pearson correlation	**-.352**
Sig. (2-tailed)	**.166**
Guess Pearson correlation	**.458**
Sig. (2-tailed)	**.065**
Metacognitive Pearson correlation	**.331**
Sig. (2-tailed)	**.194**
Nonsense Pearson correlation	**.306**
Sig. (2-tailed)	**.233**
Option-focused Pearson correlation	**- .093**
Sig. (2-tailed)	**.724**
Partial comp Pearson correlation	**-.232**
Sig. (2-tailed)	**.370**
Résumé Pearson correlation	**-.626***
Sig. (2-tailed)	**.007**
Word-bound Pearson correlation	**.067**
Sig. (2-tailed)	**.797**
Vague Pearson correlation	**-.381**
Sig. (2-tailed)	**.131**

**** = correlation is significant at the 0.02 level (2-tailed)**
*** = correlation is significant at the 0.05 level (2-tailed)**

TABLE 27 Guesses correlated with item measure

	Estimated Item Measure
Number of cases of guessing	
Pearson Correlation	.458
Sig. (2-tailed)	.065
All guesses	
Pearson Correlation	.549*
Sig. (2-tailed)	.023
Proportion of correct guesses	
Pearson Correlation	-.646**
Sig. (2-tailed)	.005

**** = correlation is significant at the 0.02 level (2-tailed)**
** = correlation is significant at the 0.05 level (2-tailed)**

TABLE 30 Elimination correlated with item measure

	Estimated Item Measure
Number of cases of elimination	
Pearson Correlation	-.066
Sig. (2-tailed)	.802
Proportion of correct elimination	
Pearson correlation	-.873**
Sig. (2-tailed)	.000

**** = correlation is significant at the 0.02 level (2-tailed)**
** = correlation is significant at the 0.05 level (2-tailed)**

TABLE 31 Number of different metacognitive responses for the different items

Response types/ Item	SIT	2LIST	LOG	OPT	UNC	CER	??	TOT
1	3	5	5	17	5	2	1	38
2	-	4	2	2	5	2	-	15
3	3	5	3	4	11	-	2	28
4	6	2	3	6	9	-	1	27
5	-	3	3	5	3	-	-	14
6	1	2	6	5	8	1	-	23
7	2	4	2	3	5	1	1	18
8	2	2	4	14	5	2	2	31
9	2	3	6	5	9	1	1	27
10	3	1	20	5	12	2	-	43
11	1	4	5	1	6	4	-	21
25	6	3	2	4	8	5	1	29
27	8	6	11	1	4	1	-	31
27	8	2	13	1	9	3	1	37
28	3	2	17	-	5	6	2	35
29	2	7	11	2	1	6	-	29
30	1	2	5	5	3	3	1	20
TOT	51	57	118	80	108	39	13	466

TABLE 33 Correlations between the number of changes of options with item measure for the 17 items

	Estimated Item Measure
Total nr of option changes	
Pearson Correlation	.225
Sig. (2-tailed)	.385
Distractor to Key	
Pearson Correlation	-.289
Sig. (2-tailed)	.261
Distractor to Distractor	
Pearson Correlation	.890**
Sig. (2-tailed)	.000
Key to Distractor	
Pearson Correlation	.132
Sig. (2-tailed)	.615

**** = correlation is significant at the 0.02 level (2-tailed)**
*** = correlation is significant at the 0.05 level (2-tailed)**

TABLE 34 The different types of metacognitive responses correlated with the item measure

		Estimated Item measure
Meta + Option-focused	Pearson Correlation	.122
	Sig. (2-tailed)	.640
Meta + Uncertainty	Pearson Correlation	.521*
	Sig. (2-tailed)	.032
Meta + Situation	Pearson Correlation	.472
	Sig. (2-tailed)	.056
Meta + 2 listening	Pearson Correlation	-.214
	Sig. (2-tailed)	.410
Meta + Logic	Pearson Correlation	-.036
	Sig. (2-tailed)	.890
Meta + Certainty	Pearson Correlation	-.136
	Sig. (2-tailed)	.602
Meta + Don’t know	Pearson Correlation	.063
	Sig. (2-tailed)	.811

**** = correlation is significant at the 0.02 level (2-tailed)**
*** = correlation is significant at the 0.05 level (2-tailed)**

APPENDIX 5

Common European Framework of Reference for Languages: Learning, Teaching, Assessment. CUP, Cambridge: General Descriptors

Proficient User	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
Independent User	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics, which are familiar, or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

http://www.coe.int/T/DG4/Portfolio/?L=E&M=/documents_intro/Data_bank_descriptors.html

JYVÄSKYLÄ STUDIES IN HUMANITIES

1 Kostiainen, Emma, Viestintä ammattiosaamisen ulottuvuutena. - Communication as a dimension of vocational competence. 305 p. Summary 4 p. 2003.

2 Seppälä, Antti, Todellisuutta kuvaamassa – todellisuutta tuottamassa. Työ ja koti television ja vähän radionkin uutisissa. - Describing reality – producing reality. Discourses of work and home in television and on a small scale in radio news. 211 p. Summary 3 p. 2003.

3 Gerlander, Maija, Jännitteet lääkärin ja potilaan välisessä viestintäsuhteessa. - Tensions in the doctor-patient communication and relationship. 228 p. Summary 6 p. 2003.

4 Lehikoinen, Taisto, Religious media theory - Understanding mediated faith and christian applications of modern media. - Uskonnollinen mediateoria: Modernin median kristilliset sovellukset. 341 p. Summary 5 p. 2003.

5 Jarva, Vesa, Venäläisperäisyys ja ekspressiivisyys suomen murteiden sanastossa. - Russian influence and expressivity in the lexicon of Finnish dialects. 215 p. 6 p. 2003.

6 Uskali, Turo, "Älä kirjoita itseäsi ulos" Suomalaisen Moskovan-kirjeenvaihtajuuden alkutaival 1957–1975. - "Do not write yourself out" The beginning of the Finnish Moscow-correspondency in 1957–1975. 484 p. Summary 4 p. 2003.

7 Valkonen, Tarja, Puheviestintätaitojen arviointi. Näkökulmia lukioikäisten esiintymis- ja ryhmätaitoihin. - Assessing speech communication skills. Perspectives on presentation and group communication skills among upper secondary school students. 310 p. Summary 7 p. 2003.

8 Tampere, Kaja, Public relations in a transition society 1989-2002. Using a stakeholder approach in organisational communications and relation analyses. 137 p. 2003.

9 Eerola, Tuomas, The dynamics of musical expectancy. Cross-cultural and statistical approaches to melodic expectations. - Musiikillisten odotusten tarkastelu kulttuurienvälisten vertailujen ja tilastollisten mallien avulla. 84 p. (277 p.) Yhteenveto 2 p. 2003.

10 Paananen, Pirkko, Monta polkua musiikkiin. Tonaalisen musiikin perusrakenteiden kehittyminen musiikin tuottamis- ja improvisaatiotehtävissä ikävuosina 6–11. - Many paths to music. The development of basic structures of tonal music in music production and improvisation at the age of 6–11 years. 235 p. Summary 4 p. 2003.

11 Laaksamo, Jouko, Musiikillisten karakterien metamorfoosi. Transformaatio- ja metamorfoosiprosessit Usko Meriläisen tuotannossa vuosina 1963–86. - "Metamorphosis of musical characters". Transformation and metamorphosis processes in the works of Usko Meriläinen during 1963–86. 307 p. Summary 3 p. 2004.

12 Rautio, Riitta, *Fortspinnungstypus* Revisited. Schemata and prototypical features in J. S. Bach's Minor-Key Cantata Aria Introductions. - Uusi katsaus kehitysmuotoon. Skeemat ja prototyyppiset piirteet J. S. Bachin kantaattien molliaarioiden alkusoitoissa. 238 p. Yhteenveto 3 p. 2004.

13 Mäntylä, Katja, Idioms and language users: the effect of the characteristics of idioms on their recognition and interpretation by native and non-native speakers of English. - Idiomien ominaisuuksien vaikutus englannin idiomien ymmärtämiseen ja tulkintaan syntyperäisten ja suomea äidinkielenään puhuvien näkökulmasta. 239 p. Yhteenveto 3 p. 2004.

14 Mikkonen, Yrjö, On conceptualization of music. Applying systemic approach to musicological concepts, with practical examples of music theory and analysis. - Musiikin käsitteellistämisestä. Systeemisen tarkastelutavan soveltaminen musikologisiin käsitteisiin sekä käytännön esimerkkejä musiikin teoriasta ja analyysistä. 294 p. Yhteenveto 10 p. 2004.

15 Holm, Jan-Markus, Virtual violin in the digital domain. Physical modeling and model-based sound synthesis of violin and its interactive application in virtual environment. - Virtuaalinen viulu digitaalisella alueella. Viulun fysikaalinen mallintaminen ja mallipohjainen äänisynteesi sekä sen vuorovaikutteinen soveltaminen virtuaalitodellisuus ympäristössä. 74 p. (123 p.) Yhteenveto 1 p. 2004.

16 Kemp, Chris, Towards the holistic interpretation of musical genre classification. - Kohti musiikin genreluokituksen kokonaisvaltaista tulkintaa. 302 p. Yhteenveto 1 p. 2004.

17 Leinonen, Kari, Finlandssvenskt sje-, tje- och s-ljud i kontrastiv belysning. 274 p. Yhteenveto 4 p. 2004.

18 Mäkinen, Eeva, Pianisti cembalistina. Cembalotekniikka cembalonsoittoa aloittavan pianistin ongelmana. - The Pianist as cembalist. Adapting to harpsichord technique as a problem for pianists beginning on the harpsichord. 189 p. Summary 4 p. 2004.

19 Kinnunen, Mauri, Herätysliike kahden kulttuurin rajalla. Lestadiolaisuus Karjalassa 1870-1939. - The Conviction on the boundary of two cultures. Laestadianism in Karelia in 1870-1939. 591 p. Summary 9 p. 2004.

20 Лилия Сиберг, "БЕЛЫЕ ЛИЛИИ". ГЕНЕЗИС Ф ИНСКОГО МИФА В БОЛГАРИИ. РОЛЬ РУССКОГО ФЕННОИЛЬСТВА. ФИНСКО-БОЛГАРСКИЕ КОНТ АКТЫ И ПОСРЕДНИКИ С КОНЦА XIX ДО КО НЦА XX ВЕКА. 284 с. - "Belye lilii". Genezis finskogo mifa v Bolgarii. Rol' russkogo fennoil'stva. Finsko-bolgarskie kontakty i posredniki s konca XIX do konca XX veka. 284 p. Yhteenveto 2 p. 2004.

JYVÄSKYLÄ STUDIES IN HUMANITIES

21 Fuchs, Bertold, Phonetische Aspekte einer Didaktik der Finnischen Gebärdensprache als Fremdsprache. - Suomalainen viittomakieli vieraana kielenä. Didaktinen fonetiikka. 476 p. Yhteenveto 14 p. 2004.

22 Jääskeläinen, Petri, Instrumentatiivisuus ja nykysuomen verbinjohto. Semanttinen tutkimus. - Instrumentality and verb derivation in Finnish. A semantic study. 504 p. Summary 5 p. 2004.

23 Mertanen Tomi, Kahdentoista markan kapina? Vuoden 1956 yleislakko Suomessa. - A Rebellion for Twelve Marks? – The General Strike of 1956 in Finland. 399 p. Summary 10 p. 2004.

24 Malherbe, Jean-Yves, L'œuvre de fiction en prose de Marcel Thiry : une lecture d'inaboutissements. 353 p. Yhteenveto 1 p. 2004.

25 Kuhna, Matti, Kahden maailman välissä. Marko Tapion *Arktinen hysteria* Väinö Linnan haastajana. - Between two worlds. Marko Tapio's Arktinen hysteria as a challenger to Väinö Linna. 307p. Summary 2 p. 2004.

26 Valtonen, Heli, Minäkuvat, arvot ja mentaliteetit. Tutkimus 1900-luvun alussa syntyneiden toimihenkilönaisten omaelämäkerroista. - Self-images, values and mentalities. An autobiographical study of white collar women in twentieth century Finland. 272 p. Summary 6 p. 2004.

27 Pusztai, Bertalan, Religious tourists. Constructing authentic experiences in late modern hungarian catholicism. - Uskontoturistit. Autenttisen elämyksen rakentaminen myöhäismodernissa unkarilaisessa katolisuudessa. 256 p. Yhteenveto 9 p. Summary in Hungarian 9 p. 2004.

28 Pääjoki, Tarja, Taide kulttuurisena kohtaamispaikkana taidekavatuksessa. - The arts as a place of cultural encounters in arts education. 125 p. Summary 3 p. 2004.

29 Juppi, Pirita, "Keitä me olemme? Mitä me haluamme?" Eläinoikeusliike määrittelykamppailun, marginalisoinnin ja moraalisen paniikin kohteena suomalaisessa sanomalehdistössä. - "Who are we? What do we want?" The animal rights movement as an object of discursive struggle, marginalization and moral panic in Finnish newspapers. 315 p. Summary 6 p. 2004.

30 Holmberg, Jukka, Etusivun politiikkaa. Yhteiskunnallisten toimijoiden representointi suomalaisissa sanomalehtiuutisissa 1987–2003. - Front page politics. Representation of societal actors in Finnish newspapers' news articles in 1987-2003. 291 p. Summary 2 p. 2004.

31 Lagerblom, Kimmo, Kaukana Kainuussa, valtaväylän varrella. Etnologinen tutkimus Kontiomäen rautatieläisyhteisön elinkaaresta 1950 – 1972. - Far, far away, nearby a main passage. An ethnological study of the life spans of Kontiomäki railtown 1950 – 1972. 407 p. Summary 2 p. 2004.

32 Hakamäki, Leena, Scaffolded assistance provided by an EFL teacher during whole-class interaction. - Vieraan kielen opettajan antama oikea-aikainen tuki luokkahuoneessa. 331 p. Yhteenveto 7 p. 2005.

33 Viergutz, Gudrun, Beiträge zur Geschichte des Musikunterrichts an den Gelehrtenschulen der östlichen Ostseeregion im 16. und 17. Jahrhundert. - Latinankoulujen musiikinopetuksen historiasta itäisen Itämeren rannikkokaupungeissa 1500- ja 1600-luvuilla. 211 p. Yhteenveto 9 p. 2005.

34 Nikula, Kaisu, Zur Umsetzung deutscher Lyrik in finnische Musik am Beispiel Rainer Maria Rilke und Einojuhani Rautavaara. - Saksalainen runous suomalaisessa musiikissa, esimerkkinä Rainer Maria Rilke ja Einojuhani Rautavaara. 304 p. Yhteenveto 6 p. 2005.

35 Syvänen, Kari, Vastatunteiden dynamiikka musiikkiterapiassa. - Counter emotions dynamics in music therapy. 186 p. Summary 4 p. 2005.

36 Eloranta, Jari & Ojala, Jari (eds), East-West trade and the cold war. 235 p. 2005.

37 Hiltunen, Kaisa, Images of time, thought and emotions: Narration and the spectator's experience in Krzysztof Kieslowski 's late fiction films. - Ajan, ajattelun ja tunteiden kuvia. Kerronta ja katsojan kokemus Krzysztof Kieslowskin myöhäisfiktiossa. 203 p. Yhteenveto 5 p. 2005.

38 Ahonen, Kalevi, From sugar triangle to cotton triangle. Trade and shipping between America and Baltic Russia, 1783-1860. 572 p. Yhteenveto 9 p. 2005.

39 Utriainen, Jaana, A gestalt music analysis. Philosophical theory, method, and analysis of Iegor Reznikoff's compositions. - Hahmoperustainen musiikkianalyysi. Hahmofilosofinen teoria, metodi ja musiikkianalyysi Iégor Reznikoffin sävellyksistä. 222 p. Yhteenveto 3 p. 2005.

40 Murtorinne, Annamari, *Tuskan hauskaa*! Tavoitteena tiedostava kirjoittaminen. Kirjoittamisprosessi peruskoulun yhdeksännellä luokalla. - Painfully fun! Towards reflective writing process. 338 p. 2005.

41 Tunturi, Anna-Riitta, Der Pikareske Roman als Katalysator in Geschichtlichen Abläufen. Erzählerische Kommunikationsmodelle in *Das Leben des Lazarillo von Tormes*, bei Thomas Mann und in EinigenFinnischen Romanen. 183 p. 2005.

42 Luoma-aho, Vilma, Faith-holders as Social Capital of Finnish Public Organisations. - Luottojoukot – Suomalaisten julkisten organisaatioiden sosiaalista pääomaa. 368 p. Yhteenveto 8 p. 2005.

JYVÄSKYLÄ STUDIES IN HUMANITIES

43 PENTTINEN, ESA MARTTI, Kielioppi virheiden varjossa. Kielitiedon merkitys lukion saksan kieliopin opetuksessa. - Grammar in the shadow of mistakes. The role of linguistic knowledge in general upper secondary school German grammar instruction. 153 p. Summary 2 p. Zusammenfassung 3 p. 2005.

44 KAIVAPALU, ANNEKATRIN, Lähdekieli kielenoppimisen apuna. - Contribution of L1 to foreign language acquisition. 348 p. Summary 7 p. 2005.

45 SALAVUO, MIIKKA,Verkkoavusteinen opiskelu yliopiston musiikkikasvatuksen opiskelukulttuurissa - Network-assisted learning in the learning culture of university music education. 317 p. Summary 5 p. 2005.

46 MAIJALA, JUHA, Maaseutuyhteisön kriisi-1930-luvun pula ja pakkohuutokaupat paikallisena ilmiönä Kalajokilaaksossa. - Agricultural society in crisis – the depression of the 1930s and compulsory sales as a local phenomenon in the basin of the Kalajoki-river. 242 p. Summary 4 p. 2005.

47 JOUHKI, JUKKA, Imagining the Other. Orientalism and occidentalism in Tamil-European relations in South India. -Tulkintoja Toiseudesta. Orientalismi ja oksidentalismi tamileiden ja eurooppalaisten välisissä suhteissa Etelä-Intiassa. 233 p. Yhteenveto 2 p. 2006.

48 LEHTO, KEIJO, Aatteista arkeen. Suomalaisten seitsenpäiväisten sanomalehtien linjapapereiden synty ja muutos 1971–2005. - From ideologies to everyday life. Editorial principles of Finnish newspapers, 1971–2005. 499 p. Summary 3 p. 2006.

49 VALTONEN, HANNU, Tavallisesta kuriositeetiksi. Kahden Keski-Suomen Ilmailumuseon Messerschmitt Bf 109 -lentokoneen museoarvo. - From Commonplace to curiosity – The Museum value of two Messerschmitt Bf 109 -aircraft at the Central Finland Aviation Museum. 104 p. 2006.

50 KALLINEN, KARI, Towards a comprehensive theory of musical emotions. A multi-dimensional research approach and some empirical findings. - Kohti kokonaisvaltaista teoriaa musiikillisista emootioista. Moniulotteinen tutkimuslähestymistapa ja empiirisiä havaintoja. 71 p. (200 p.) Yhteenveto 2 p. 2006.

51 ISKANIUS, SANNA, Venäjänkielisten maahanmuuttajaopiskelijoiden kieli-identiteetti. - Language and identity of Russian-speaking students in Finland. 264 p. Summary 5 p. Реферат 6 с. 2006.

52 HEINÄNEN, SEIJA, Käsityö – taide – teollisuus. Näkemyksiä käsityöstä taideteollisuuteen 1900-luvun alun ammatti- ja aikakauslehdissä. - Craft – Art – Industry: From craft to industrial art in the views of magazines and trade publications of the early 20th Century. 403 p. Summary 7 p. 2006.

53 KAIVAPALU, ANNEKATRIN & PRUULI, KÜLVI (eds), Lähivertailuja 17. - Close comparisons. 254 p. 2006.

54 ALATALO, PIRJO, Directive functions in intra-corporate cross-border email interaction. - Direktiiviset funktiot monikansallisen yrityksen englanninkielisessä sisäisessä sähköpostiviestinnässä. 471 p. Yhteenveto 3 p. 2006.

55 KISANTAL, TAMÁS, „…egy tömegmészárlásról mi értelmes dolgot lehetne elmondani?" Az ábrázolásmód mint történelemkoncepció a holokauszt-irodalomban. - "...there is nothing intelligent to say about a massacre". The representational method as a conception of history in the holocaust-literature. 203 p. Summary 4 p. 2006.

56 MATIKAINEN, SATU, Great Britain, British Jews, and the international protection of Romanian Jews, 1900-1914: A study of Jewish diplomacy and minority rights. - Britannia, Britannian juutalaiset ja Romanian juutalaisten kansainvälinen suojelu, 1900–1914: Tutkimus juutalaisesta diplomatiasta ja vähemmistöoikeuksista. 237 p. Yhteenveto 7 p. 2006.

57 HÄNNINEN, KIRSI, Visiosta toimintaan. Museoiden ympäristökasvatus sosiokulttuurisena jatkumona, säätelymekanismina ja innovatiivisena viestintänä. - From vision to action. Environmental education in museums as a socio-cultural continuum, regulating mechanism, and as innovative communication 278 p. Summary 6 p. 2006.

58 JOENSUU, SANNA, Kaksi kuvaa työntekijästä. Sisäisen viestinnän opit ja postmoderni näkökulma. - Two images of an employee; internal communication doctrines from a postmodern perspective. 225 p. Summary 9 p. 2006.

59 KOSKIMÄKI, JOUNI, Happiness is… a good transcription - Reconsidering the Beatles sheet music publications. - Onni on… hyvä transkriptio – Beatles-nuottijulkaisut uudelleen arvioituna. 55 p. (320 p. + CD). Yhteenveto 2 p. 2006.

60 HIETAHARJU, MIKKO, Valokuvan voi repiä. Valokuvan rakenne-elementit, käyttöympäristöt sekä valokuvatulkinnan syntyminen. - Tearing a photograph. Compositional elements, contexts and the birth of the interpretation. 255 p. Summary 5 p. 2006.

61 JÄMSÄNEN, AULI, Matrikkelitaiteilijaksi valikoituminen. Suomen Kuvaamataiteilijat -hakuteoksen (1943) kriteerit. - Prerequisites for being listed in a biographical encyclopedia criteria for the Finnish Artists Encyclopedia of 1943. 285 p. Summary 4 p. 2006.

62 HOKKANEN, MARKKU, Quests for Health in Colonial Society. Scottish missionaries and medical culture in the Northern Malawi region, 1875-1930. 519 p. Yhteenveto 9 p. 2006.

JYVÄSKYLÄ STUDIES IN HUMANITIES

63 Ruuskanen, Esa, Viholliskuviin ja viranomaisiin vetoamalla vaiennetut työväentalot. Kuinka Pohjois-Savon Lapuan liike sai nimismiehet ja maaherran sulkemaan 59 kommunistista työväentaloa Pohjois-Savossa vuosina 1930–1932. - The workers' halls closed by scare-mongering and the use of special powers by the authorities. 248 p. Summary 5 p. 2006.

64 Vardja, Merike, Tegelaskategooriad ja tegelase kujutamise vahendid Väinö Linna romaanis “Tundmatu sõdur”. - Character categories and the means of character representation in Väinö Linna’s Novel *The Unknown Soldier*. 208 p. Summary 3 p. 2006.

65 Takáts, József, Módszertani berek. Írások az irodalomtörténet-írásról. - The Grove of Methodology. Writings on Literary Historiography. 164 p. Summary 3 p. 2006.

66 Mikkola, Leena, Tuen merkitykset potilaan ja hoitajan vuorovaikutuksessa. - Meanings of social support in patient-nurse interaction. 260 p. Summary 3 p. 2006.

67 Saarikallio, Suvi, Music as mood regulation in adolescence. - Musiikki nuorten tunteiden säätelynä. 46 p. (119 p.) Yhteenveto 2 p. 2007.

68 Hujanen, Erkki, Lukijakunnan rajamailla. Sanomalehden muuttuvat merkitykset arjessa. - On the fringes of readership. The changing meanings of newspaper in everyday life. 296 p. Summary 4 p. 2007.

69 Tuokko, Eeva, Mille tasolle perusopetuksen englannin opiskelussa päästään? Perusopetuksen päättövaiheen kansallisen arvioinnin 1999 eurooppalaisen viitekehyksen taitotasoihin linkitetyt tulokset. - What level do pupils reach in English at the end of the comprehensive school? National assessment results linked to the common European framework. 338 p. Summary 7 p. Sammanfattning 1 p. Tiivistelmä 1 p. 2007.

70 Tuikka, Timo, ”Kekkosen konstit”. Urho Kekkosen historia- ja politiikkakäsitykset teoriasta käytäntöön 1933–1981. - ”Kekkonen´s way”. Urho Kekkonen’s conceptions of history and politics from theory to practice, 1933–1981 413 p. Summary 3 p. 2007.

71 Humanistista kirjoa. 145 s. 2007.

72 Nieminen, Lea, A complex case: a morphosyntactic approach to complexity in early child language. 296 p. Tiivistelmä 7 p. 2007.

73 Torvelainen, Päivi, Kaksivuotiaiden lasten fonologisen kehityksen variaatio. Puheen ymmärrettävyyden sekä sananmuotojen tavoittelun ja tuottamisen tarkastelu. - Variation in phonological development of two-year-old Finnish children. A study of speech intelligibility and attempting and production of words. 220 p. Summary 10 p. 2007.

74 Siitonen, Marko, Social interaction in online multiplayer communities. - Vuorovaikutus verkkopeliyhteisöissä. 235 p. Yhteenveto 5 p. 2007.

75 Stjernvall-Järvi, Birgitta, Kartanoarkkitehtuuri osana Tandefelt-suvun elämäntapaa. - Manor house architecture as part of the Tandefelt family´s lifestyle. 231 p. 2007.

76 Sulkunen, Sari, Text authenticity in international reading literacy assessment. Focusing on PISA 2000. - Tekstien autenttisuus kansainvälisissä lukutaidon arviointitutkimuksissa: PISA 2000. 227 p. Tiivistelmä 6 p. 2007.

77 Kőszeghy, Péter, Magyar Alkibiadés. Balassi Bálint élete. - The Hungarian Alcibiades. The life of Bálint Balass. 270 p. Summary 6 p. 2007.

78 Mikkonen, Simo, State composers and the red courtiers - Music, ideology, and politics in the Soviet 1930s - Valtion säveltäjiä ja punaisia hoviherroja. Musiikki, ideologia ja politiikka 1930-luvun Neuvostoliitossa. 336 p. Yhteenveto 4 p. 2007.

79 Sivunen, Anu, Vuorovaikutus, viestintäteknologia ja identifioituminen hajautetuissa tiimeissä. - Social interaction, communication technology and identification in virtual teams. 251 p. Summary 6 p. 2007.

80 Lappi, Tiina-Riitta, Neuvottelu tilan tulkinnoista. Etnologinen tutkimus sosiaalisen ja materiaalisen ympäristön vuorovaikutuksesta jyväskyläläisissä kaupunkipuhunnoissa. - Negotiating urban spatiality. An ethnological study on the interplay of social and material environment in urban narrations on Jyväskylä. 231 p. Summary 4 p. 2007.

81 Huhtamäki, Ulla, ”Heittäydy vapauteen”. Avantgarde ja Kauko Lehtisen taiteen murros 1961–1965. - ”Fling yourself into freedom!” The Avant-Garde and the artistic transition of Kauko Lehtinen over the period 1961–1965. 287 p. Summary 4 p. 2007.

82 Kela, Maria, *Jumalan kasvot* suomeksi. Metaforisaatio ja erään uskonnollisen ilmauksen synty. - God’s face in Finnish. Metaphorisation and the emergence of a religious expression. 275 p. Summary 5 p. 2007.

83 Saarinen, Taina, Quality on the move. Discursive construction of higher education policy from the perspective of quality. - Laatu liikkeessä. Korkeakoulupolitiikan diskursiivinen rakentuminen laadun näkökulmasta. 90 p. (176 p.) Yhteenveto 4 p. 2007.

84 Mäkilä, Kimmo, Tuhoa, tehoa ja tuhlausta. Helsingin Sanomien ja New York Timesin ydinaseuutisoinnin tarkastelua diskurssianalyyttisesta näkökulmasta 1945–1998.

- ”Powerful, Useful and Wasteful”. Discourses of Nuclear Weapons in the New York Times and Helsingin Sanomat 1945–1998. 337 p. Summary 7 p. 2007.

85 Kantanen, Helena, Stakeholder dialogue and regional engagement in the context of higher education. - Yliopistojen sidosryhmävuoropuhelu ja alueellinen sitoutuminen. 209 p. Yhteenveto 8 p. 2007.

86 Almonkari, Merja, Jännittäminen opiskelun puheviestintätilanteissa. - Social anxiety in study-related communication situations. 204 p. Summary 4 p. 2007.

87 Valentini, Chiara, Promoting the European Union. Comparative analysis of EU communication strategies in Finland and in Italy. 159 p. (282 p.) 2008.

88 Pulkkinen, Hannu, Uutisten arkkitehtuuri - Sanomalehden ulkoasun rakenteiden järjestys ja jousto. - The Architecture of news. Order and flexibility of newspaper design structures. 280 p. Yhteenveto 5 p. 2008.

89 Meriläinen, Merja, Monenlaiset oppijat englanninkielisessä kielikylpyopetuksessa - rakennusaineita opetusjärjestelyjen tueksi. - Diverse Children in English Immersion: Tools for Supporting Teaching Arrangements. 197 p. 2008.

90 Vares, Mari, The question of Western Hungary/Burgenland, 1918-1923. A territorial question in the context of national and international policy. - Länsi-Unkarin/Burgenlandin kysymys 1918–1923. Aluekysymys kansallisen ja kansainvälisen politiikan kontekstissa. 328 p. Yhteenveto 8 p. 2008.

91 Ala-Ruona, Esa, Alkuarviointi kliinisenä käytäntönä psyykkisesti oireilevien asiakkaiden musiikkiterapiassa – strategioita, menetelmiä ja apukeinoja. – Initial assessment as a clinical procedure in music therapy of clients with mental health problems – strategies, methods and tools. 155 p. 2008.

92 Oravala, Juha, Kohti elokuvallista ajattelua. Virtuaalisen todellisen ontologia Gilles Deleuzen ja Jean-Luc Godardin elokuvakäsityksissä. - Towards cinematic thinking. The ontology of the virtually real in Gilles Deleuze’s and Jean-Luc Godard’s conceptions of cinema. 184 p. Summary 6 p. 2008.

93 Kecskeméti, István, Papyruksesta megabitteihin. Arkisto- ja valokuvakokoelmien konservoinnin prosessin hallinta. - From papyrus to megabytes: Conservation management of archival and photographic collections. 277 p. 2008.

94 Suni, Minna, Toista kieltä vuorovaikutuksessa. Kielellisten resurssien jakaminen toisen kielen omaksumisen alkuvaiheessa. - Second language in interaction: sharing linguistic resources in the early stage of second language acquisition. 251 p. Summary 9 p. 2008.

95 N. Pál, József, Modernség, progresszió, Ady Endre és az Ady–Rákosi vita. Egy konfliktusos eszmetörténeti pozíció természete és következményei. 203 p. Summary 3 p. 2008.

96 Bartis, Imre, „Az igazság ismérve az, hogy igaz”. Etika és nemzeti identitás Sütő András Anyám könnyű álmot ígér című művében és annak recepciójában. 173 p. Summary 4 p. 2008.

97 Ranta-Meyer, Tuire, Nulla dies sine linea. Avauksia Erkki Melartinin vaikutteisiin, verkostoihin ja vastaanottoon henkilö- ja reseptiohistoriallisena tutkimuksena. - *Nulla dies sine linea*: A biographical and reception-historical approach to Finnish composer Erkki Melartin. 68 p. Summary 6 p. 2008.

98 Koivisto, Keijo, Itsenäisen Suomen kanta-aliupseeriston synty, koulutus, rekrytointitausta ja palvelusehdot. - The rise, education, the background of recruitment and conditions of service of the non-commissioned officers in independent Finland. 300 p. Summary 7 p. 2008.

99 Kiss, Miklós, Between narrative and cognitive approaches. Film theory of non-linearity applied to Hungarian movies. 198 p. 2008.

100 Ruusunen, Aimo, Todeksi uskottua. Kansandemokraattinen Neuvostoliitto-journalismi rajapinnan tulkkina vuosina1964–1973. - Believed to be true. Reporting on the USSR as interpretation of a boundary surface in pro-communist partisan journalism 1964–1973. 311 p. Summary 4 p. 2008.

101 Härmälä, Marita, Riittääkö *Ett ögonblick* näytöksi merkonomilta edellytetystä kielitaidosta? Kielitaidon arviointi aikuisten näyttötutkinnoissa. – Is *Ett ögonblick* a sufficient demonstration of the language skills required in the qualification of business and administration? Language assessment in competence-based qualifications for adults. 318 p. Summary 4 p. 2008.

102 Coelho, Jacques, The vision of the cyclops. From painting to video ways of seeing in the 20th century and through the eyes of Man Ray. 538 p. 2008.

103 Brewis, Kielo, Stress in the multi-ethnic customer contacts of the Finnish civil servants: Developing critical pragmatic intercultural professionals. – Stressin kokemus suomalaisten viranomaisten monietnisissä asiakaskontakteissa: kriittis-pragmaattisen kulttuurienvälisen ammattitaidon kehittäminen. 299 p. Yhteenveto 4 p. 2008.

104 Belik, Zhanna, The Peshekhonovs’ Workshop: The Heritage in Icon Painting. 239 p. [Russian]. Summary 7 p. 2008.

105 Moilanen, Laura-Kristiina, Talonpoikaisuus, säädyllisyys ja suomalaisuus 1800- ja 1900-lukujen vaihteen suomenkielisen proosan kertomana. – Peasant values, estate society and the Finnish in late nineteenth- and early

and early twentieth-century narrative literature. 208 p. Summary 3 p. 2008.

106 Päärnilä, Ossi, Hengen hehkusta tietostrategioihin. Jyväskylän yliopiston humanistisen tiedekunnan viisi vuosikymmentä. 110 p. 2008.

107 Kangasniemi, Jukka, Yksinäisyyden kokemisen avainkomponentit Yleisradion tekstitelevision Nuorten palstan kirjoituksissa. - The key components of the experience of loneliness on the Finnish Broadcasting Company's (YLE) teletext forum for adolescents. 388 p. 2008.

108 Gajdó, Tamás, Színháztörténeti metszetek a 19. század végétől a 20. század közepéig. - Segments of theatre history from the end of the 19th century to the middle of the 20th century. 246 p. Summary 2 p. 2008.

109 Catani, Johanna, Yritystapahtuma kontekstina ja kulttuurisena kokemuksena. - Corporate event as context and cultural experience. 140 p. Summary 3 p. 2008.

110 Mahlamäki-Kaistinen, Riikka, Mätänevän velhon taidejulistus. Intertekstuaalisen ja -figuraalisen aineiston asema Apollinairen L'Enchanteur pourrissant teoksen tematiikassa ja symboliikassa. - Pamphlet of the rotten sorcerer. The themes and symbols that intertextuality and interfigurality raise in Apollinaire's prose work L'Enchanteur pourrissant. 235 p. Résumé 4 p. 2008.

111 Pietilä, Jyrki, Kirjoitus, juttu, tekstielementti. Suomalainen sanomalehtijournalismi juttutyyppien kehityksen valossa printtimedian vuosina 1771-2000. - Written Item, Story, Text Element. Finnish print journalism in the light of the development of journalistic genres during the period 1771-2000. 779 p. Summary 2 p. 2008.

112 Saukko, Päivi, Musiikkiterapian tavoitteet lapsen kuntoutusprosessissa. - The goals of music therapy in the child's rehabilitation process. 215 p. Summary 2 p. 2008.

113 Lassila-Merisalo, Maria, Faktan ja fiktion rajamailla. Kaunokirjallisen journalismin poetiikka suomalaisissa aikakauslehdissä. - On the borderline of fact and fiction. The poetics of literary journalism in Finnish magazines. 238 p. Summary 3 p. 2009.

114 Knuutinen, Ulla, Kulttuurihistoriallisten materiaalien menneisyys ja tulevaisuus. Konservoinnin materiaalitutkimuksen heritologiset funktiot. - The heritological functions of materials research of conservation. 157 p. (208 p.) 2009.

115 Niiranen, Susanna, «Miroir de mérite». Valeurs sociales, rôles et image de la femme dans les textes médiévaux des *trobairitz*. - "Arvokkuuden peili". Sosiaaliset arvot, roolit ja naiskuva keskiaikaisissa *trobairitz*-teksteissä. 267 p. Yhteenveto 4 p. 2009.

116 Aro, Mari, Speakers and doers. Polyphony and agency in children's beliefs about language learning. - Puhujat ja tekijät. Polyfonia ja agentiivisuus lasten kielenoppimiskäsityksissä. 184 p. Yhteenveto 5 p. 2009.

117 Jantunen, Tommi, Tavu ja lause. Tutkimuksia kahden sekventiaalisen perusyksikön olemuksesta suomalaisessa viittomakielessä. - Syllable and sentence. Studies on the nature of two sequential basic units in Finnish Sign Language. 64 p. 2009.

118 Särkkä, Timo, Hobson's Imperialism. A Study in Late-Victorian political thought. - J. A. Hobsonin imperialismi. 211 p. Yhteenveto 11 p. 2009.

119 Laihonen, Petteri, Language ideologies in the Romanian Banat. Analysis of interviews and academic writings among the Hungarians and Germans. 51 p. (180 p) Yhteenveto 3 p. 2009.

120 Mátyás, Emese, Sprachlernspiele im DaF-Unterricht. Einblick in die Spielpraxis des finnischen und ungarischen Deutsch-als-Fremdsprache-Unterrichts in der gymnasialen Oberstufe sowie in die subjektiven Theorien der Lehrenden über den Einsatz von Sprachlernspielen. 399 p. 2009.

121 Paraczky, Ágnes, Näkeekö taitava muusikko sen minkä kuulee? Melodiadiktaatin ongelmat suomalaisessa ja unkarilaisessa taidemusiikin ammattikoulutuksessa. - Do accomplished musicians see what they hear? 164 p. Magyar nyelvü összefoglaló 15 p. Summary 4 p. 2009.

122 Elomaa, Eeva, Oppikirja eläköön! Teoreettisia ja käytännön näkökohtia kielten oppimateriaalien uudistamiseen. - Cheers to the textbook! Theoretical and practical considerations on enchancing foreign language textbook design. 307 p. Zusammanfassung 1 p. 2009.

123 Helle, Anna, Jäljet sanoissa. Jälkistrukturalistisen kirjallisuuskäsityksen tulo 1980-luvun Suomeen. - Traces in the words. The advent of the poststructuralist conception of literature to Finland in the 1980s. 272 p. Summary 2 p. 2009.

124 Pimiä, Tenho Ilari, Tähtäin idässä. Suomalainen sukukansojen tutkimus toisessa maailmansodassa. - Setting sights on East Karelia: Finnish ethnology during the Second World War. 275 p. Summary 2 p. 2009.

125 Vuorio, Kaija, Sanoma, lähettäjä, kulttuuri. Lehdistöhistorian tutkimustraditiot Suomessa ja median rakennemuutos. - Message, sender, culture. Traditions of research into the history of the press in Finland and structural change in the media. 107 p. 2009.

126 Bene, Adrián Egyén és közösség. Jean-Paul Sartre *Critique de la raison dialectique* című műve a magyar recepció tükrében. - Individual and community. Jean-Paul Sartre's

Critique of dialectical reason in the mirror of the Hungarian reception. 230 p. Summary 5 p. 2009.

127 Drake, Merja, Terveysviestinnän kipupisteitä. Terveystiedon tuottajat ja hankkijat Internetissä. - At the interstices of health communication. Producers and seekers of health information on the Internet. 206 p. Summary 9 p. 2009.

128 Rouhiainen-Neunhäuserer, Maijastiina, Johtajan vuorovaikutusosaaminen ja sen kehittyminen. Johtamisen viestintähaasteet tietoperustaisessa organisaatiossa. - The interpersonal communication competence of leaders and its development. Leadership communication challenges in a knowledge-based organization. 215 p. Summary 9 p. 2009.

129 Vaarala, Heidi, Oudosta omaksi. Miten suomenoppijat keskustelevat nykynovellista? - From strange to familiar: how do learners of Finnish discuss the modern short story? 317 p. Summary 10 p. 2009.

130 Marjanen, Kaarina, The Belly-Button Chord. Connections of pre-and postnatal music education with early mother-child interaction. - Napasointu. Pre- ja postnataalin musiikkikasvatuksen ja varhaisen äiti-vauva -vuorovaikutuksen yhteydet. 189 p. Yhteenveto 4 p. 2009.

131 Bőhm, Gábor, Önéletírás, emlékezet, elbeszélés. Az emlékező próza hermeneutikai aspektusai az önéletírás-kutatás újabb eredményei tükrében. - Autobiography, remembrance, narrative. The hermeneutical aspects of the literature of remembrance in the mirror of recent research on autobiography. 171 p. Summary 5 p. 2009.

132 Leppänen, Sirpa, Pitkänen-Huhta, Anne, Nikula, Tarja, Kytölä, Samu, Törmäkangas, Timo, Nissinen, Kari, Kääntä, Leila, Virkkula, Tiina, Laitinen, Mikko, Pahta, Päivi, Koskela, Heidi, Lähdesmäki, Salla & Jousmäki, Henna, Kansallinen kyselytutkimus englannin kielestä Suomessa: Käyttö, merkitys ja asenteet. - National survey on the English language in Finland: Uses, meanings and attitudes. 365 p. 2009.

133 Heikkinen, Olli, Äänitemoodi. Äänite musiikillisessa kommunikaatiossa. - Recording Mode. Recordings in Musical Communication. 149 p. 2010.

134 Lähdesmäki, Tuuli (Ed.), Gender, Nation, Narration. Critical Readings of Cultural Phenomena. 105 p. 2010.

135 Mikkonen, Inka, "Olen sitä mieltä, että". Lukiolaisten yleisönosastotekstien rakenne ja argumentointi. - "In my opinion…" Structure and argumentation of letters to the editor written by upper secondary school students. 242 p. Summary 7 p. 2010.

136 Nieminen, Tommi, Lajien synty. Tekstilaji kielitieteen semioottisessa metateoriassa. - Origin of genres: Genre in the semiotic metatheory of linguistics. 303 p. Summary 6 p. 2010.

137 Kääntä, Leila, Teacher turn allocation and repair practices in classroom interaction. A multisemiotic perspective. - Opettajan vuoronanto- ja korjauskäytänteet luokkahuonevuorovaikutuksessa: multisemioottinen näkökulma. 295 p. Yhteenveto 4 p. 2010. HUOM: vain verkkoversiona.

138 Saarimäki, Pasi, Naimisen normit, käytännöt ja konfliktit. Esiaviollinen ja aviollinen seksuaalisuus 1800-luvun lopun keskisuomalaisella maaseudulla. - The norms, practices and conflicts of sex and marriage. Premarital and marital sexual activity in rural Central Finland in the late nineteenth century. 275 p. Summary 12 p. 2010.

139 Kuuva, Sari, Symbol, Munch and creativity: Metabolism of visual symbols. - Symboli, Munch ja luovuus – Visuaalisten symboleiden metabolismi. 296 p. Yhteenveto 4 p. 2010.

140 Skaniakos, Terhi, Discoursing Finnish rock. Articulations of identities in the Saimaa-Ilmiö rock documentary. - Suomi-rockin diskursseja. Identiteettien artikulaatioita Saimaa-ilmiö rockdokumenttielokuvassa. 229 p. 2010.

141 Kauppinen, Merja, Lukemisen linjaukset – lukutaito ja sen opetus perusopetuksen äidinkielen ja kirjallisuuden opetussuunnitelmissa. - Literacy delineated – reading literacy and its instruction in the curricula for the mother tongue in basic education. 338 p. Summary 8 p. 2010.

142 Pekkola, Mika, Prophet of radicalism. Erich Fromm and the figurative constitution of the crisis of modernity. - Radikalismin profeetta. Erich Fromm ja modernisaation kriisin figuratiivinen rakentuminen. 271 p. Yhteenveto 2 p. 2010.

143 Kokkonen, Lotta, Pakolaisten vuorovaikutussuhteet. Keski-Suomeen muuttaneiden pakolaisten kokemuksia vuorovaikutussuhteistaan ja kiinnittymisestään uuteen sosiaaliseen ympäristöön. - Interpersonal relationships of refugees in Central Finland: perceptions of relationship development and attachment to a new social environment. 260 p. Summary 8 p. 2010.

144 Kananen, Heli Kaarina, Kontrolloitu sopeutuminen. Ortodoksinen siirtoväki sotien jälkeisessä Ylä-Savossa (1946-1959). - Controlled integration: Displaced orthodox Finns in postwar upper Savo (1946–1959). 318 p. Summary 4 p. 2010.

JYVÄSKYLÄ STUDIES IN HUMANITIES

145 Nissi, Riikka, Totuuden jäljillä. Tekstin tulkinta nuorten aikuisten raamattupiirikeskusteluissa. – In search of the truth. Text interpretation in young adults’ Bible study conversations. 351 p. Summary 5 p. 2010.

146 Lilja, Niina, Ongelmista oppimiseen. Toisen aloittamat korjausjaksot kakkoskielisessä keskustelussa. – Other-initiated repair sequences in Finnish second language interactions. 336 p. Summary 8 p. 2010.

147 Váradi, Ildikó, A parasztpolgárosodás „finn útja”. Kodolányi János finnországi tevékenysége és finn útirajzai. – The “Finnish Way” of Peasant-Bourgeoization. János Kodolányi’s Activity in Finland and His Travelogues on Finland. 182 p. Summary 3 p. 2010.

148 Hankala, Mari, Sanomalehdellä aktiiviseksi kansalaiseksi? Näkökulmia nuorten sanomalehtien lukijuuteen ja koulun sanomalehtiopetukseen. – Active citizenship through newspapers? Perspectives on young people´s newspaper readership and on the use of newspapers in education. 222 p. Summary 5 p. 2011.

149 Salminen, Elina, Monta kuvaa menneisyydestä. Etnologinen tutkimus museokokoelmien yksityisyydestä ja julkisuudesta. – Images of the Past. An ethnological study of the privacy and publicity of museum collections. 226 p. Summary 5 p. 2011. HUOM: vain verkkoversiona.

150 Järvi, Ulla, Media terveyden lähteillä. Miten sairaus ja terveys rakentuvat 2000-luvun mediassa. – Media forces and health sources. Study of sickness and health in the media. 209 p. Summary 3 p. 2011.

151 Ullakonoja, Riikka, Da. Eto vopros! Prosodic development of Finnish students´ read-aloud Russian during study in Russia. – Suomalaisten opiskelijoiden lukupuhunnan prosodinen kehittyminen vaihto-opiskelujakson aikana Venäjällä. 159 p. (208 p.) Summary 5 p. 2011.

152 Marita Vos, Ragnhild Lund, Zvi Reich and Halliki Harro-Loit (Eds), Developing a Crisis Communication Scorecard. Outcomes of an International Research Project 2008-2011 (Ref.). 340 p. 2011.

153 Punkanen, Marko, Improvisational music therapy and perception of emotions in music by people with depression. 60 p. (94 p.) Yhteenveto 1 p. 2011.

154 Di Rosario, Giovanna, Electronic poetry. Understanding poetry in the digital environment. – Elektroninen runous. Miten runous ymmärretään digitaalisessa ympäristössä? 327 p. Tiivistelmä 1 p. 2011.

155 Tuuri, Kai, Hearing Gestures: Vocalisations as embodied projections of intentionality in designing non-speech sounds for communicative functions. – Puheakteissa kehollisesti välittyvä intentionaalisuus apuna ei-kielellisesti viestivien käyttöliittymä-äänien suunnittelussa. 50 p. (200 p.) Yhteenveto 2 p. 2011.

156 Martikainen, Jari, Käsitettävä taidehistoria. Kuvalähtöinen malli taidehistorian opetukseen kuvallisen ilmaisun ammatillisessa perustutkinnossa. – Grasping art history. A picture-based model for teaching art history in the vocational basic degree programme in visual arts. 359 p. Summary 10 p. 2011.

157 Hakanen, Marko, Vallan verkostoissa. Per Brahe ja hänen klienttinsä 1600-luvun Ruotsin valtakunnassa. – Networks of Power: Per Brahe and His Clients in the Sixteenth-Century Swedish Empire. 216 p. Summary 6 p. 2011.

158 Lindström, Tuija Elina, Pedagogisia merkityksiä koulun musiikintunneilla perusopetuksen yläluokkien oppilaiden näkökulmasta. – Pedagogical Meanings in Music Education from the Viewpoint of Students of Junior High Grades 7-9. 215 p. 2011.

159 Anckar, Joanna, Assessing foreign language listening comprehension by means of the multiple-choice format: processes and products. – Vieraan kielen kuullun ymmärtämistaidon mittaaminen monivalintatehtävien avulla: prosesseja ja tuloksia. 308 p. Tiivistelmä 2 p. 2011.